

Yiyun Su

25 Ave at Port Imperial West New York NJ USA 07093 | 815-517-4203|

yiyun.su@rutgers.edu | <https://linkedin.com/in/yiyun-su> | <https://scholar.google.com> | <https://github.com/suiyun>

RESEARCH INTERESTS

Data-driven AI Systems | Trustworthy Machine Learning | Vision-Language Models | Agentic AI | Knowledge Grounding | Software Engineering | Distributed Systems | Financial Information Systems

EDUCATION

Rutgers University

Master of Information Technology GPA 3.6

New Brunswick, New Jersey, USA

8/2020- 12/2021

Northern Illinois University

Bachelor of Science in Information Systems GPA 3.6

DeKalb, Illinois, USA

08/2016- 05/2018

Beijing Technology and Business University

Bachelor of Engineering in Automation Systems GPA 3.4

Beijing, China

08/2014- 07/2016

RESEARCH EXPERIENCE

Multi-Agent LLM Simulation: Web3 + Open Claw Agent Village

01/2026-present

Target MISQ Journal submission: Multi-Agent Systems · Emergent Behavior · LLM Evaluation

- Led experimental design for a multi-agent LLM simulation system operating in a Web3-inspired virtual environment; defined agent role specifications and interaction constraints to model realistic autonomous behavior
- Crafted structured inter-agent communication prompts to enable goal-directed collaboration and negotiation across heterogeneous agent types
- Established quantitative evaluation metrics to assess emergent behavior, task completion rates, and coordination efficiency across simulation runs

GraphRAGQuery: Scalable Knowledge Ingestion for Agentic Text-to-SQL

07/2025-01/2026

Enterprise RAG System · Knowledge Graph · Prompt Engineering

- Proposed and implemented a scalable knowledge ingestion framework targeting enterprise RAG systems for agentic Text-to-SQL applications; grounded in taxonomy findings from ICIC 2026 survey
- Engineered multi-stage prompt pipelines to decompose complex natural language queries into structured SQL, incorporating schema linking and intermediate reasoning steps
- Designed graph-based knowledge representation to capture relational structure across enterprise data sources, improving query grounding accuracy over flat retrieval baselines

Agentic-SQL Text-to-SQL Survey Frameworks and Benchmarking Analysis

09/2025 – present

Accepted · ICIC 2026 · [link]

- Led a comprehensive survey and benchmark analysis of Text-to-SQL frameworks; systematically categorized methods across parsing strategies, schema linking, and execution accuracy metrics
- Proposed a taxonomy of evaluation dimensions and identified key gaps in cross-domain generalization and agentic query decomposition, motivating follow-on system design in GraphRAGQuery

Typed Strategy Programs for Bayesian Exploration in LLM-Based Agents

09/2025 – 05/2026

NeurIPS 2026 Submission · StratEx · Program-Level Thompson Sampling · LLM Agent Scaffold

- Conducted hyperparameter sensitivity analysis across exploration rate, intrinsic reward weight, and controller learning rate; validated robustness of StratEx on SWE-bench Lite, WebArena, and ToolBench-IR benchmarks
- Performed frontier backbone swap experiments on Claude Opus 4.7 and GPT-5 under controlled same-scaffold protocol; confirmed that program-level exploration gains (+5–7 pp over action-level Thompson Sampling) persist across model scales
- **Publication (Under Review, NeurIPS 2026):** StratEx: Typed Strategy Programs for Bayesian Exploration in LLM-Based Agents

MM: Merging Models via Important Neurons

08/2025 – 02/2026

ACM Multimedia '26 Submission · Model Merging · Neuron Pruning · Router-based Fusion

- Designed and executed comprehensive CV benchmark evaluation across **10 datasets** (MNIST, EuroSAT, CIFAR-10, GTSRB, DTD, RESISC, and others) using **ViT-Base-Patch16-224**; reproduced fine-tuned baselines and merging competitors including Weight Averaging, Task-Arithmetic, Ties-Merging, and Twin-Merging under unified evaluation protocol
- Conducted scalability analysis experiments measuring normalized accuracy and storage footprint as the number of merged tasks scales from 2 to 10 (Figure 4); confirmed that MM maintains stable performance and lower memory overhead compared to Twin-Merging as task count increases
- **Publication (Under Review, ACM Multimedia '26):** MM: Merging Models via Important Neurons

Brain Lesion Segmentation and Automated Diagnosis via Hilbert-Mamba Enhanced SAM2

04/2025 – 08/2025

Multi-Framework Research Initiative · SAM-GPT · CenterMamba-SAM · Hilbert-VLM

- Conducted systematic literature review and comparative study of SAM2, MedSAM, and Mamba-based segmentation architectures; validated on BraTS2021, FCD2023, and ISLES2022 benchmarks
- **Publication :** Bridging the Perception-Cognition Gap: Re-engineering SAM2 with Hilbert-Mamba for Robust VLM-based Medical Diagnosis
- **Publication in BIBM 2025:** CenterMamba-SAM: Center-Prioritized Scanning and Temporal Prototypes for Brain Lesion Segmentation

PROFESSIONAL EXPERIENCE

- Axon** Seattle, Washington
Software Development Engineer 06/2025 - present
- Built internal AI-powered developer tooling integrating **Splunk** and **Grafana** via **MCP**, enabling automated log analysis, anomaly detection, and code review assistance across engineering teams
 - Architected **zero-downtime** migration of distributed messaging infrastructure (**Kafka**, **Temporal**, **Cassandra**) across 12 microservices from Azure to AWS to enable AI-native capabilities, including direct integration with **AWS Bedrock** for **LLM-powered workflows**; achieved **60%** reduction in deployment time for mission-critical evidence management systems. (evidence.com)
 - Standardized **CI/CD** pipeline templates in **GitHub Actions** to accelerate delivery cycles and enforce test coverage and code quality.
- Amazon Web Service** New York City, New York, USA
Software Development Engineer 12/2022 - 05/2025
- Led a Financial Reporting Service for AWS Bedrock Marketplace Commerce platform to support **automated payments billing systems** for 1P and third party (3P) seller models using **Java Spring** framework and make internal data connect with 2 new Restful API using AWS services: **SNS**, **SQS**, **Lambda**, **DynamoDB**.
 - Created new data sources web connectors for data crawling data site on AWS Bedrock - Knowledge Base ((launch link)- Structured RAG Application for AI Agent application infra, with new API and **AWS SDK**, including **rate limiting**, fetching, and **ingestion scaling** to support multiple data sources for RAG, increased web crawling **4x throughput** to support **40** companies with **Authn/Authz**.
 - **Launched Amazon Managed Blockchain Query Service (June 2023)**: Established operational readiness system by ensuring system readiness and robust security using **CloudFormation** infrastructure as code and **CloudWatch** for monitoring dashboard. ((press link)
 - **Initiated Data Pipelines for AWS Managed Blockchain Query System: Launched Jan 2024** Handled Bitcoin and Ethereum historical backfill data (> 500 TB data) from **S3** to **DynamoDB** using **Step Function** and **Spark UDF** in **Glue**, confirming data validation results as 100
 - **Designed and Implemented new APIs** to enable Bitcoin UTXO feature and list batch and streaming transactional data for blockchain addresses in the current and given time frame using **API Gateway**.
- Visa Technology** Remote Austin, Texas, USA
Systems Analyst - Production Reliability Engineer 02/2022 - 12/2022
- Supported **Visa Installment** Service as contactless payment as a **Dev Ops Engineer** by using **Kubernetes** containers and deploy weekly using **Jenkins CI/CD**, ensured the release work by **SQL** using **DataGrip**
 - Deployed **release** activities in **Unix Servers** for both Sandbox and Production environment, switching traffic for globally different servers in **RedShift** to support **10M+** customers who using contactless payments like Apple Pay.
- RingCentral** Fort Lauderdale, Florida, USA
Software Engineer Intern 06/2021 - 09/2021
- Established Glip-Express-Errors library from **design** to **implementation** using **Typescript** and improved the HTTP error handling process by **50%** efficiency
 - Initiated the **npm** package with Node.js environment setup and **YAML** file and **Typescript ESLint** in **AWS sandbox** environment
 - Conducted **test driven development** in **Jest**, submitted and received code reviews to **GitLab** code base and **Jenkins CI/CD** pipeline
 - Collaborated with the upgrade of the **OpenAPI** validator used in versioned router and created versioned routes for Reactions
 - Headed a disruptive technology - **Turn cold call to smart call** project and won second place at the final company's intern project
- Computer Science Department, RPI** Troy, New York, USA
Teaching Assistant 08/2018 - 05/2019
- Assisted course about (CS1100) Computer Science I for **Python Programming** and (CS2600) Principles of Software using **Java**
 - Hosted 4 hour lab session every week for 100+ students, received 93% grade on student feedback

SKILLS

- **Distributed Systems**: Kafka, Cassandra, DynamoDB, Redis
- **Cloud-native Infrastructure**:: AWS (Lambda, S3, ECS, Step Functions) Data Engineering: Spark, Glue, ETL pipelines
- **Coding Languages**: Java, Python, TypeScript
- **ML AI tools**: PyTorch, TensorFlow, HuggingFace Transformers, LangChain LlamaIndex, Pinecone