

# SHICHENG LIU

shicheng@cs.stanford.edu ◇ 773-236-6337

## EDUCATION

---

Stanford University

**Ph.D. in Computer Science**

Jun. 2027 (*expected*)

Natural Language Processing Group, 3rd year

The University of Chicago

**(Honors) B.S. Computer Science** *with a specialization in Computer Systems*

Jun. 2022

**(Honors) B.S. Mathematics**

**Minor in Physics**

Cumulative GPA: 3.985/4.000 (*summa cum laude*)

California Institute of Technology

Quarter-long Exchange. *Exchange major: Computer Science*

Sept. - Dec. 2021

Exchange GPA: 4.1/4.3

## RESEARCH INTEREST

---

*Research areas: Natural Language Processing, Computer Systems, Programming Languages*

I focus on real-life, practical NLP problems, often drawing perspectives from computer systems and programming languages. My recent research focuses on knowledge agents with LLMs, aiming to enable domain-independent approaches that effectively retrieve and navigate different sources of knowledge, including structured, unstructured, and hybrid (combination of structured and unstructured data) sources.

## PUBLICATIONS

---

### **SPINACH: SPARQL-Based Information Navigation for Challenging Real-World Questions**

**Shicheng Liu\***, Sina J. Semnani\*, Harold Triedman, Jialiang Xu, Isaac Dan Zhao, Monica S. Lam.

*Findings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024)*

TL;DR: Introduces a novel Knowledge-Base Question Answering (KBQA) dataset and agent. The SPINACH dataset introduces challenging, expert-annotated questions from Wikidata's request query forum. The SPINACH agent, mimicking how human experts write SPARQL queries, outperforms previous models across multiple KBQA datasets. [The deployed SPINACH agent](#) and the [online chat interface](#) have since been actively used by the Wikidata community.

[\[code\]](#) [\[video\]](#) [\[blog\]](#)

### **SUQL: Conversational Search over Structured and Unstructured Data with Large Language Models**

**Shicheng Liu**, Jialiang Xu, Wesley Tjangnaka, Sina J. Semnani, Chen Jie Yu, Monica S. Lam.

*Findings of the North American Chapter of the Association for Computational Linguistics: NAACL 2024*

TL;DR: Introduces the first conversational agent capable of accessing both structured and unstructured data from large knowledge corpora using a new language called SUQL (Structured and Unstructured Query Language), which extends SQL with free-text capabilities based on retrievers and LLMs. SUQL compiler performs important optimizations to power hybrid queries. Experiments on HybridQA and user studies on Yelp ([Online Demo](#)) show that a SUQL-based agent outperforms strong baselines

[\[code\]](#) [\[video\]](#)

## Fine-tuned LLMs Know More, Hallucinate Less with Few-Shot Sequence-to-Sequence Semantic Parsing over Wikidata

Silei Xu\*, **Shicheng Liu**\*, Theo Culhane, Elizaveta Pertseva, Meng-Hsi Wu, Sina Semnani, Monica Lam.  
*Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)*

TL;DR: Introduces WikiWebQuestions, a KBQA benchmark for Wikidata converted from the popular WebQuestionSP dataset. It presents a few-shot semantic parser based on fine-tuned version of LLaMA for Wikidata, with modified SPARQL syntax to enhance accuracy. When paired with GPT-3, the system can provide useful answers to 96% of the questions in the dev set of WikiWebQuestions.

[\[code\]](#) [\[video\]](#) [\[blog\]](#)

## Coding Reliable LLM-based Integrated Task and Knowledge Agents with GenieWorksheets

Harshit Joshi, **Shicheng Liu**, James Chen, Robert Weigle, Monica S. Lam.

*arXiv preprint, 2024/07, in submission*

TL;DR: Introduces a programmable framework for creating task and knowledge conversational agents that handle complex interactions. GenieWorksheets enables developers to program agent policies through its declarative paradigm. The compiled agent is resilient to diverse user queries and helpful with knowledge sources. It outperforms GPT-4 in execution accuracy, dialogue act accuracy, and goal completion rate, with results validated through real user studies.

## SPAGHETTI: Open-Domain Question Answering from Heterogeneous Data Sources with Retrieval and Semantic Parsing

Heidi C. Zhang, Sina J. Semnani, Farhad Ghassemi, Jialiang Xu, **Shicheng Liu**, Monica S. Lam.

*Findings of the Association for Computational Linguistics: ACL 2024*

TL;DR: Introduces SPAGHETTI: Semantic Parsing Augmented Generation for Hybrid English information from Text Tables and Infoboxes, a hybrid question-answering (QA) pipeline that utilizes information from heterogeneous knowledge sources, including knowledge base, text, tables, and infoboxes. This LLM-augmented approach achieves SOTA performance on the Compmix dataset, the most comprehensive heterogeneous open-domain QA dataset

## Automated Testing of Software that Uses Machine Learning APIs

Chengcheng Wan, **Shicheng Liu**, Sophie Xie, Yifan Liu, Henry Hoffmann, Michael Maire, Shan Lu.

*Proceedings of the 44th International Conference on Software Engineering, 2022*

## Are Machine Learning Cloud APIs Used Correctly?

Chengcheng Wan, **Shicheng Liu**, Henry Hoffmann, Michael Maire, Shan Lu.

*Proceedings of the 43th International Conference on Software Engineering, 2021*

## SELECTED HONORS & AWARDS

---

### Teachings

**Top-5% of Stanford CS Course Assistants:** Stanford CS224V

2023

### Grants & Scholarships

**2024 Brown Institute Magic Grant** (\$80,000 grant)

2024-2025

- Leading a bi-coastal collaboration between members from Stanford CS, Stanford Big Local News, and Columbia Journalism on *DataTalk: All Documents and Data, All at Once, All Verified*.

- *Project Overview*: Investigative journalism often relies on the ability to mine diverse data sets, with both structured and unstructured forms. Building on the novel programming language SUQL, this project aims to develop trustworthy conversational agents for journalists to uncover insights from hybrid data sources using natural-language queries. System available at <https://datatalk.genie.stanford.edu/>. Example of published article using our agent on [Atlanta Journal Constitution](#).

<b>College Summer Research Fellow</b> (\$5,000 grant)	2021
<b>College Research Fellow</b> (\$4,500 grant)	2020-2021
<b>Soong Ching Ling Foundation Scholarship</b> (\$12,500 scholarship)	2020
<b>Jeff Metcalf Summer Research Fellowship</b> (\$6,000 grant)	2019

#### Academic Honors

<b>Graduated Summa Cum Laude</b> , The University of Chicago	2022
<b>Outstanding Undergraduate Researcher Award</b> , Honorable Mention, CRA	2022

- [CRA website notice](#)
- Featured on [UChicago CS News](#)

<b>Elected member of Phi Beta Kappa</b> , the University of Chicago (the Beta chapter of Illinois)	2021
<b>Enrico Fermi Scholar</b> , The University of Chicago	2021
<b>Robert Maynard Hutchins Scholar</b> , The University of Chicago	2020
<b>Dean's List</b> , The University of Chicago	2018-2019, 2019-2020, 2020-2021

### SELECTED RESEARCH EXPERIENCE

<b>Researcher</b>	Jun. 2022 - Present
Stanford Open Virtual Assistant Lab, Department of Computer Science, Stanford University	
<b>Researcher</b>	Jan. 2020 - Jun. 2022
Prof. Shan Lu Research Group, Department of Computer Science, The University of Chicago	
<b>Research Intern</b>	Jun. 2019 - Oct. 2019
Laboratory for Space Research, The University of Hong Kong, China ( <i>UChicago Jeff Metcalf Intern</i> )	

### TECHNICAL SKILLS

Python, C, C++, Rust, SQL, Java/Type script, R, Julia, Haskell, Standard ML, Racket

### TEACHING EXPERIENCES

#### Stanford

**Head Course Assistant** for CS 224V Conversational Virtual Assistants with Deep Learning Fall 2023

#### University of Chicago

<b>Teaching Assistant</b> for DATA 12000 Computer Science for Data Science	Spring 2022
<b>Teaching Assistant</b> for CMSC 27200 Theory of Algorithms	Winter 2022
<b>Grader</b> for CMSC 22100 Programming Languages	Spring 2021
<b>Teaching Assistant</b> for CMSC 15100 Introduction to Computer Science I	Winter 2021

## RELEVANT COURSES

---

<b>CMSC 16100</b>	<i>(Honors)</i> Introduction to Computer Science I
<b>CMSC 15200</b>	Introduction to Computer Science II
<b>CMSC 15400</b>	Introduction to Computer System
<b>CMSC 22100</b>	Programming Languages
<b>CMSC 23000</b>	Operating Systems
<b>CMSC 23010</b>	Parallel Computing
<b>CMSC 23500</b>	Introduction to Database Systems
<b>CMSC 25700</b>	Natural Language Processing
<b>CMSC 27100</b>	Discrete Mathematics
<b>CMSC 27200</b>	Theory of Algorithms
<b>CMSC 27410</b>	<i>(Honors)</i> Combinatorics
<b>CMSC 28000</b>	Introduction to Formal Languages
<b>CS 152</b>	<i>(Caltech)</i> Introduction to Cryptography
<b>CS/CNS 171</b>	<i>(Caltech)</i> Computer Graphics Laboratory
<b>CS 144</b>	<i>(Stanford)</i> Introduction to Computer Networking
<b>CS 256</b>	<i>(Stanford)</i> Algorithmic Fairness
<b>CS 261</b>	<i>(Stanford)</i> Optimization and Algorithmic Paradigms
<b>TTIC 31020</b>	<i>(PhD-level)</i> Introduction to Machine Learning
<b>MATH 16x00</b>	<i>(Honors)</i> Calculus I-II-III
<b>MATH 20x10</b>	<i>(Accelerated)</i> Analysis in $\mathbb{R}^n$ I-II-III
<b>MATH 20250</b>	Abstract Linear Algebra
<b>MATH 23500</b>	Markov Chains, Martingales, and Brownian Motion
<b>MATH 25x00</b>	<i>(Honors)</i> Basic Algebra I-II-III
<b>MATH 26500</b>	Introduction to Riemannian Geometry
<b>MATH 27000</b>	Basic Complex Variables
<b>Ma 109A</b>	<i>(Caltech)</i> Introduction to Geometry and Topology
<b>Ma 116A</b>	<i>(Caltech)</i> Mathematical Logic and Axiomatic Set Theory
<b>PHYS 14100</b>	<i>(Honors)</i> Mechanics
<b>PHYS 14200</b>	<i>(Honors)</i> Electricity & Magnetism
<b>PHYS 14300</b>	<i>(Honors)</i> Waves, Optics, & Heat
<b>PHYS 15400</b>	Modern Physics
<b>PHYS 18500</b>	Intermediate Mechanics
<b>PHYS 23x00</b>	Quantum Mechanics I-II

## ACADEMIC REFERENCES

---

<b>Prof. Monica S. Lam</b> <i>(Ph.D. advisor)</i>	<a href="mailto:lam@cs.stanford.edu">lam@cs.stanford.edu</a>
Kleiner Perkins, Mayfield, Sequoia Capital Professor of the School of Engineering , Department of Computer Science Stanford University, Stanford, U.S.A.	
<b>Prof. Shan Lu</b> <i>(Undergrad advisor)</i>	<a href="mailto:shanlu@uchicago.edu">shanlu@uchicago.edu</a>
Professor, Department of Computer Science The University of Chicago, Chicago, U.S.A.	

*This CV is last updated on Dec. 1st, 2024*