SHICHENG LIU

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

EDUCATION

Stanford University
Ph.D. in Computer Science 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 Langauge 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, generating 1,700+ conversations.
[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 and Honolulu Civil Beat.

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

Academic Honors

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

- CRA website notice
- Featured on UChicago CS News

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

SELECTED RESEARCH EXPERIENCE

Researcher Jun. 2022 - Present
Stanford Open Virtual Assistant Lab, Department of Computer Science, Stanford University

Researcher Jan. 2020 - Jun. 2022
Prof. Shan Lu Research Group, Department of Computer Science, The University of Chicago

Research Intern Jun. 2019 - Oct. 2019
Laboratory for Space Research, The University of Hong Kong, China
(UChicago Jeff Metcalf Intern)

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

Teaching Assistant for DATA 12000 Computer Science for Data Science Spring 2022

Teaching Assistant for CMSC 27200 Theory of Algorithms Winter 2022

Grader for CMSC 22100 Programming Languages Spring 2021

Teaching Assistant for CMSC 15100 Introduction to Computer Science I Winter 2021

RELEVANT COURSES

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

ACADEMIC REFERENCES

Prof. Monica S. Lam *(Ph.D. advisor)* `lam@cs.stanford.edu`
Kleiner Perkins, Mayfield, Sequoia Capital Professor of the School of Engineering , Department of Computer Science
Stanford University, Stanford, U.S.A.

Prof. Shan Lu *(Undergrad advisor)* `shanlu@uchicago.edu`
Professor, Department of Computer Science
The University of Chicago, Chicago, U.S.A.

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