SHICHENG LIU

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

EDUCATION

Stanford University
Ph.D. in Computer Science 2026 *(expected)*
Natural Language Processing Group, 4th 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.

SELECTED RESEARCH EXPERIENCE

AI Research Intern Jun. - Sept. 2025
Meta Reality Labs (in collaboration with FAIR at Meta)

Research Assistant Jun. 2022 - Present
Stanford Open Virtual Assistant Lab & Stanford NLP Group

PUBLICATIONS

SCRIBES: Web-Scale Script-Based Semi-Structured Data Extraction with Reinforcement Learning
Shicheng Liu, Kai Sun, Lisheng Fu, Xilun Chen, Xinyuan Zhang, Zhaojiang Lin, Rulin Shao, Yue Liu, Anuj Kumar, Wen-tau Yih, Xin Luna Dong
Pre-print, under review

TL;DR: SCRIBES is a reinforcement learning framework for large-scale extraction of semi-structured web data (tables, lists, infoboxes). Instead of running costly per-page LLM inference, it generates reusable scripts for groups of structurally similar pages, using layout similarity as a reward signal. Trained on both high-quality human annotations and synthetic CommonCrawl annotations, SCRIBES outperforms baselines by 13% in script quality and improves GPT-4o's downstream QA accuracy by over 4%, making web information extraction more scalable and efficient, with more potential for complex QAs and model pre-training.

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 22,000+ 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 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.
of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025)

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. Example of published article using our agent on Atlanta Journal Constitution and Honolulu Civil Beat.

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

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

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 October 22, 2025