# Sitao Cheng

+1-805-722-6280 | sitaocheng@ucsb.edu | https://sitaocheng.github.io/

## Research Interest

I am passionate about LLM-Agents, Retrieval-augmented Generation (RAG) and Neural-Symbolic Reasoning. I have worked on knowledge-intensive reasoning, including structured (Knowledge Base, Table), unstructured (Document) and models parametric knowledge. Currently, I focus on the interplay between parametric and contextual knowledge.

## Education

- **Nanjing University** *09.2021 - 06.2024*
  *M.S. in Computer Science and Technology - Grade: 92.34 / 100.00 (Top 5%)* Nanjing, China
- **University of Electronic Science and Technology of China** *09.2017 - 06.2021*
  *B.E. in Software Engineering - GPA: 3.98 / 4.00 (Top 3)* Chengdu, China

## Publications

*Equal Contribution

**Conference paper**

[1] **Call Me When Necessary: LLMs can Efficiently and Faithfully Reason over Structured Environments**. ACL, 2024. [link]
**Sitao Cheng**, Ziyuan Zhuang, Yong Xu, Fangkai Yang, Chaoyun Zhang, Xiaoting Qin, Xiang Huang, Ling Chen, Qingwei Lin, Dongmei Zhang, Saravan Rajmohan, Qi Zhang

[2] **QueryAgent: a Reliable and Efficient Reasoning Framework with Environmental Feedback-based Self-Correction**. ACL (Oral), 2024. [link]
Xiang Huang*, **Sitao Cheng***, Shanshan Huang, Jiayu Shen, Yong Xu, Chaoyun Zhang, Yuzhong Qu

[3] **MarkQA: a Large Scale KBQA Dataset with Numerical Reasoning**. EMNLP, 2023. [link]
Xiang Huang, **Sitao Cheng**, Yuheng Bao, Shanshan Huang, Yuzhong Qu

[4] **Question Decomposition Tree for Answering Complex Questions over Knowledge Bases**. AAAI (Oral), 2023. [link]
Xiang Huang, **Sitao Cheng**, Yiheng Shu, Yuheng Bao, Yuzhong Qu

[5] **EfficientRAG: Efficient Retriever for Multi-Hop Question Answering**. EMNLP, 2024. [link]
Ziyuan Zhuang*, Zhiyang Zhang*, **Sitao Cheng**, Fangkai Yang, Jia Liu, Shujian Huang, Qingwei Lin, Saravan Rajmohan, Dongmei Zhang, Qi Zhang

**Preprints.**

[1] **Understanding the Interplay between Parametric and Contextual Knowledge for Large Language Models**. ICLR Under Review [link]
**Sitao Cheng**, Liangming Pan, Xunjian Yin, Xinyi Wang, William Yang Wang

[2] **Disentangling Memory and Reasoning Ability in Large Language Models**. ARR Under Review [link]
Mingyu Jin, Weidi Luo, **Sitao Cheng**, Xinyi Wang, Wenyue Hua, Ruixiang Tang, William Yang Wang, Yongfeng Zhang

[3] **Thread: A Logic-Based Data Organization Paradigm for How-To Question Answering with Retrieval Augmented Generation**. ICLR Under Review [link]
Kaikai An, Fangkai Yang, Liqun Li, Junting Lu, **Sitao Cheng**, Shuzheng Si, Lu Wang, Pu Zhao, Lele Cao, Qingwei Lin, Saravan Rajmohan, Dongmei Zhang, Qi Zhang, Baobao Chang

## Research Experience

- **University of California, Santa Barbara (NLP Group)** *07.2024 - Now*
  ***Advisor**: Prof. William Wang. **Role**: Visiting Research Scholar* Santa Barbara, U.S.A
  - **Topic**: Understanding how effective LLMs leverage parametric knowledge when contextual knowledge is given.
    - Description: We systematically design various relationships between the two knowledge sources: *supportive, complementary, conflicting and irrelevant*. We introduce a new dataset EchoQA across scientific, factual and commonsense knowledge, to access models ability of echoing their knowledge given contextual information.
    - Findings: LLMs consistently **suppress their own knowledge** given the context, regardless of models, knowledge types, the relations between two knowledge sources, and various levels of instructions.
    - Results: One submission on ICLR 2025.
- **Microsoft Research Asia (DKI Group)** *10.2023 - 06.2024*
  ***Advisor**: Yong Xu, Fangkai Yang, Chaoyun Zhang. **Role**: Research Intern & Mentor of Junior Interns* Beijing, China

- ◦ **Topic 1**: LLMs reasoning framework over structured environments with retrieval-augmented generation (**Readi**) or neural symbolic reasoning (**QueryAgent**).
  - ∗ Description: With large-scaled and heterogeneous structured environments (e.g. Knowledge Graphs, Tables, Databases, etc), how LLMs can reason both efficiently and faithfully? Our intuition is from humans exploration with real-world environments. We adopt LLMs to either directly maintain a reasoning path (Readi), or step-by-step build a query (QueryAgent), both incorporating pertinent information for correction.
  - ∗ Results: Two publications on ACL 2024.
- ◦ **Topic 2**: Efficient iterative retrieval with soley encoder-based models (**EfficientRAG**) and a new data organization paradigm (**THREAD**) for RAG systems.
  - ∗ Description: For better retrieval, it is crucial to model the link between the chunks. We leverage strong understanding ability of LLMs to reason the link between chunks. We design novel retrieval methods for smaller encoder-based models (EfficientRAG) and re-organize the documents (Thread), to model such link.
  - ∗ Results: One publication on EMNLP 2024. One submission on ICLR 2025.
- ◦ **Topic 3**: LLM-based Personalized Assistant with "**SurpriseMe**" interaction by Structured Knowledge Graphs.
  - ∗ Description: LLMs not only answer questions with powerful conversational capabilities, but also provide human beings with emotion and interest assistance tailored to their individual experience.
  - ∗ Results: One submission on CSCW 2025.

- **Nanjing University (Websoft Lab)** *09.2021 - 06.2024*
  ***Advisor**: Prof. Yuzhong Qu. **Role**: Student Researcher* Nanjing, China
  - ◦ **Topic 1**: Step-by-step query building (**QueryAgent**) with self-correction based on environmental feedback.
    - ∗ Description: In-context learning generates the query on one go, which is unreliable. While current incremental query-building method suffers from hallucination problems, we introduce a functional toolset with environmental feedback and a zero-shot correction method for both reliability and efficiency.
    - ∗ Results: One publication on ACL 2024.
  - ◦ **Topic 2**: A KBQA benchmark (**MarkQA**) requiring both multi-hop and numerical reasoning ability.
    - ∗ Description: We propose NR-KBQA to challenge both reasoning ability over knowledge bases. We build a dataset (MarkQA), scaling automatically to 32k from a small number of seeds. We design PyQL query, a function toolset able to seamless SPARQL conversion, as symbolic reasoning steps, alleviating labeling burden.
    - ∗ Results: One publication on EMNLP 2023.
  - ◦ **Topic 3**: A question decomposition method (**QDT**) for better multi-hop reasoning over knowledge bases.
    - ∗ Description: We propose a serializable tree-based structure (QDT) to represent complex questions, which can sufficiently split questions with complex structures. We also propose a two-staged generative based method (Clue-Decipher) to ease the uncontrollable nature of generative LMs.
    - ∗ Results: One publication on AAAI 2023.
- **Ant Group** *06.2023 - 09.2023*
  ***Advisor**: Xiaoyin Chu (Digitization Group). **Role**: Research Intern* Hangzhou, China
  - ◦ **Topic**: Adopt LLMs to build knowledge graph based on long documents. Denoise and expand the text chunks for better multi-hop question answering.
    - ∗ Description: In real-world scenarios, language models tend to hallucination with long context. We adopt LLMs to process documents into triple sets and adopt multi-chain reasoning in RAG systems.

## Honors and Awards

- **ACL 2024 Volunteer**
- **ARR Reviewer**
- **Outstanding Student of Sichuan Province**
- **Outstanding Graduate Student Award** – NJU, UESTC
- **Academic Scholarship * 5** – NJU, UESTC
- **MCM/ICM H Prize**

## Skills

- **Professional Skills:** Popular NLP models (LLM applications, Transformers, attention mechanism, etc.), Pytorch, C++, LaTex, Python, SQL
- **Languages:** TOEFL 105, CET-4 CET-6 Excellent
- **Interests:** Body building (over 6x body weight in the Big 3) , Basketball (member of department team), Swim
- **Social Service:** I serve as a personal assistant for a disabled senior impressionist artist in UC Santa Barbara.