Sitao Cheng

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

RESEARCH INTEREST

I am passionate about LLM-Agents, Retrieval-augmented Generaltion (RAG) and Neural-Symbolic Reasoning. I have experience on reasoning over real-world environments (e.g. Knowledge Base, Tables as structured and documents as unstructured environments). Currently, I focus on the understanding of LLMs utilization of knowledge.

EDUCATION

- **Nanjing University** — *09.2021 - 06.2024*
  M.S. in Computer Science and Technology - Grade: 92.35/100.00 (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**.
Sitao Cheng, Ziyuan Zhuang, Yong Xu, Fangkai Yang, Chaoyun Zhang, Xiaoting Qin, Xiang Huang, Ling Chen, Qingwei Lin, Dongmei Zhang, Saravan Rajmohan, Qi Zhang
ACL (Findings), 2024. [link]

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

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

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

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

Preprints.

[1] **Understanding the Interplay between Parametric and Contextual Knowledge for Large Language Models**.
Sitao Cheng, Liangming Pan, Xunjian Yin, Xinyi Wang, William Yang Wang
ICLR (Under Review), 2025.

[2] **Thread: A Logic-Based Data Organization Paradigm for How-To Question Answering with Retrieval Augmented Generation**.
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
ICLR (Under Review), 2025. [link]

RESEARCH EXPERIENCE

- **University of California, Santa Barbara** — *07.2024 - Now*
  Advisor: Prof. William Wang (UCSB NLP Group). Role: Visiting Research Scholar — Santa Barbara, U.S.A
  ◦ **Topic**: Understanding how effective LLMs leverage parametric knowledge when contextual knowledge is given.
    ∗ Description: We identify 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 knowledge types or its relations between two knowledge sources.
    ∗ Results: One submission on ICLR 2025.

- **Microsoft Research Asia** — *10.2023 - 06.2024*
  Advisor: Yong Xu, Fangkai yang, Chaoyun Zhang (DKI Group). Role: Research Intern — Beijing, China

- **Topic 1**: LLMs reasoning 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 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 fine-tune smaller encoder-based models (EfficientRAG) or 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 "Surprising" interaction by Structured Knowledge Bases.
    - ∗ 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 CHI 2024.

- **Nanjing University**   *09.2021 - 06.2024*
  Advisor: Prof. Yuzhong Qu (Websoft Lab). **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 correction method for better efficiency and reliability.
        - ∗ Results: One publication on ACL 2024.
    - **Topic 2**: A KBQA benchmark (MarkQA) requiring both multi-hop reasoning 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, which can be converted into SPARQL, 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 Question Decomposition Tree (QDT) structure to represent natural language 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 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, we 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 the documents into triple sets and adopt multi-chain reasoning for better results.

HONORS AND AWARDS

- **Outstanding Graduate Student Award**   *06.2024*
  NJU
- **Outstanding Student of Sichuan Province**   *06.2021*
  UESTC
- **Outstanding Graduate Student Award**   *06.2021*
  UESTC
- **First Prize Academic Scholarship * 3**   *2021-2025*
  UESTC, NJU
- **Second Prize Academic Scholarship * 2**   *2021-2025*
  NJU
- **MCM/ICM H Prize**   *06.2021*
  MCM/ICM

SKILLS

- **Professional Skills:** Common NLP models (LLM applications, Transformers, attention mechanism, etc.), Pytorch, C++, LaTex, Python, SQL
- **Languages:** Good English speaking and listening skills (TOEFL 106, CET-4 CET-6 Excellent)
- **Interests:** Body building (over 6x body weight in the Big 3) , Basketball (member of department team, Swimming