StructLLM: The first workshop on Learning Large Language Models with Structured Data

Kaize Ding[◇], Yue Zhao[⋆], Yu Zhang[◦], Yu Wang[†], Xiusi Chen[∗], Qi Zhu[‡], Yuan Luo[◇]
[◇]Northwestern Univeristy, [⋆]University of Southern California [◦]Texas A&M University
[†]University of Oregon, [∗]University of llinois Urbana-Champaign, [‡]Amazon

1 Workshop Topic

Large language models (LLMs) have transformed AI by learning from massive collections of *unstructured* text, demonstrating impressive capabilities in language understanding, reasoning, and few-shot learning. However, much of the world's information is not in free text but in *structured* formats such as tables, time series, graphs, and multi-modal semi-structured records. Such structured data is ubiquitous across applications in different high-impact domains, ranging from predictive analytics in finance and molecular generation in biochemistry to medical diagnosis in healthcare. Yet, structured data remains underexplored in the era of artificial general intelligence, with current LLMs poorly equipped to process, integrate, or reason over it.

To date, research on structured data has largely advanced in separate silos, with limited cross-community dialogue on how to tailor LLMs to these modalities. This workshop seeks to bring together researchers and practitioners across disciplines to chart a unified agenda for LLMs and structured data. By fostering collaboration, establishing benchmarks, and inspiring new methods, we aim to accelerate progress toward LLMs that fully harness the richness of structured data. We welcome contributions on theory, data, modeling, evaluation, applications, and governance. By bridging structured data research with LLM advances, this workshop seeks to lay the groundwork for the next generation of foundation models that can seamlessly learn from and reason over the world's rich and diverse structured data sources. Specifically, we highlight the following topics:

LLMs for Structured Data. LLMs have shown remarkable capabilities in predictive, reasoning, and generation tasks. Recently, researchers have explored leveraging LLMs for structured data and shown promising results. However, this field is still in its infancy and challenges remain, including (1) the interpretability of LLM-based approaches, (2) their ability to handle large-scale structured datasets efficiently, and (3) their reliability in processing multi-modal structured data. We welcome contributions on LLMs for different structured data sources, including tabular data, time series, relational databases, and semi-structured records, etc.

Data-Centric AI for LLMs on Structured Data. Advancing LLMs for structured data requires innovations in how diverse data modalities are represented, integrated, and leveraged during training. Challenges include handling heterogeneous formats, missing or noisy values, schema understanding, and reasoning over multi-modal structured inputs. We welcome contributions on pre-training strategies, data augmentation, retrieval-augmented generation, structured reasoning, and domain adaptation that improve LLM efficiency, robustness, and generalization across structured data.

Benchmarking LLMs on Structured Data. While efforts have been made to develop benchmarks for different structured data modalities, new efforts are required to evaluate LLMs on structured data comprehensively along different dimensions, such as inference throughput, memory usage, scalability, and data memorization. Another challenge exists in evaluating a foundation model where contamination of the training data exists. Mitigating contamination is particularly hard for data-hungry foundation models in the data-scarce structured data domains. We welcome contributions of novel benchmarks evaluating FMs for structured data.

Applications of LLMs for Structured Data. The development of LLMs for structured data can transform numerous applications, from climate modeling and fraud detection to molecular design and medical diagnosis. Deploying these models in real-world settings requires addressing key challenges such as model reliability and data privacy. This workshop seeks contributions: (1) showcasing novel applications of LLMs in structured data domains, (2) overcoming challenges, such as scaling and inference throughput for existing applications, and (3) demonstrating domain-specific innovation. We also welcome discussions on ethical considerations, fairness, and bias mitigation in different applications and domains.

2 Organizational Information

2.1 Format of the Workshop

Our full-day workshop will consist of the following main components: (1) *invited keynotes* from experts in the field of LLMs for structured data coming *from both industry and academia* to create a synergistic atmosphere and to stimulate collaborations, (2) contributed *research oral talks* selected from the set of accepted works, (3) *panel discussion* that will be composed of our keynote speakers given their expertise in this domain, and 4) contributed *poster session* to allow accepted works to present and socialize. We are in contact with the following tentative speakers and panelists and will finalize the arrangement with them upon the acceptance of our workshop.

Jiawei Han, Michael Aiken Chair Professor at University of Illinois Urbana-Champaign

Heng Ji, Professor at University of Illinois Urbana-Champaign

Derek Zhiyuan Cheng, Principal Scientist & Research Director at Google DeepMind

James Caverlee, Professor at Texas A&M University

Bryan Perrozi, Tech Lead Manager at Google Research

Sercan Arik, Research Manager at Google

Luna Dong, Principal Scientist at Meta

Shirui Pan, Professor at Griffith University

Qingsong Wen, Head of AI at Squirrel AI Learning

The workshop will try to have corporate sponsorships for funding the speakers, diversity and inclusion efforts, and best paper awards.

2.2 Hybrid/Virtual Delivery

The workshop will feature live-streamed and recorded sessions to facilitate the needs of remote attendees and speakers. Remote and in-person attendees will have equal presentation opportunity. The workshop will accommodate remote poster presentation by providing assistance in poster printing, setup, and video QA. The organizers will provide clear instructions and timely reminders for remote and in-person attendees and speakers.

2.3 Specicial Requirements and Restrictions

The preferred venue of the workshop is ACL 2026 due to visa-related travel restrictions of key organizers and the time availabilities of certain invited speakers.

The workshop requires standard equipments (i.e. projector, microphones, poster boards, camera, etc.) for oral presentation, poster session, and panel discussion.

3 Diversity and Inclusion

The organizers are committed to enhance the diversity and inclusion of the workshop by:

Promoting Academic Diversity. We welcome a broad spectrum of contributions, including position papers, theoretical advances, empirical studies, systems reports, datasets and benchmarks, and case studies from both academia and industry. Submissions may draw on real-world deployments and non-public resources (e.g., proprietary data, prototype datasets, production logs), provided that results are responsibly summarized, anonymized where necessary, and fully compliant with privacy regulations and ACL policy requirements. We also explicitly encourage negative results, ablation studies, and lessons learned that illuminate failure modes, practical trade-offs, and reproducibility challenges. Special priority will be given to work that engages with underrepresented user groups, languages, and regions.

Diversifying Representation. Our invited speakers, panelists, organizers, and program committee members will reflect a wide range of perspectives. We are committed to ensuring the inclusion of individuals from underrepresented groups in computational linguistics, women in science, and LGBTQ+ communities. We will also maintain a balanced mix of junior and senior researchers, encourage participation from both academia and industry, and promote international representation spanning both developed and developing regions.

Diversifying Participation. To lower barriers to engagement, we will provide registration subsidies for selected in-person and online attendees, with preference given to individuals from marginalized and underrepresented groups, those facing financial constraints, and participants affected by visa restrictions. Throughout the workshop, we will uphold ACL's Code of Ethics and Anti-Harassment Policy to ensure a safe, welcoming, and inclusive environment for all attendees.

A Organizing Committee

Kaize Ding is an Assistant Professor in the Department of Statistics and Data Science at Northwestern University. Before joining Northwestern, he obtained his Ph.D. degree in Computer Science at Arizona State University in 2023. His research interests are generally in data mining, machine learning, and natural language processing, with a particular focus on graph machine learning, data-efficient learning, and reliable AI. His work has been recognized with several prestigious awards and honors, including the AAAI New Faculty Highlights, SDM Best Posters Award, etc.

Yue Zhao is an Assistant Professor of Computer Science at the University of Southern California and a faculty member of the USC Machine Learning Center. His research focuses on trustworthy and efficient AI, including anomaly and out-of-distribution detection, graph learning, and open-source ML systems. He has served as Workflow Co-Chair for KDD 2023 and is co-organizing the AI for Financial Fraud Detection & Prevention workshop at ICAIF 2025. He also leads the FORTIS Lab and develops community resources such as PyOD, PyGOD, TDC, and TrustLLM.

Yu Zhang is an Assistant Professor of Computer Science & Engineering at Texas A&M University. He received his Ph.D. degree from the University of Illinois at Urbana-Champaign, advised by Prof. Jiawei Han. His research interests include NLP for science and scientific research, as well as text mining with graphs and structured knowledge. Yu is the recipient of the ACM SIGKDD Dissertation Award Runner-Up, the UIUC Dissertation Completion Fellowship, and the Yunni & Maxine Pao Memorial Fellowship. He has served as a (Senior) Area Chair for ACL, EMNLP, NeurIPS, and KDD.

Yu Wang is an Assistant Professor at the Department of Computer Science at the University of Oregon, who conducts research in topology-informed graph machine learning and knowledge retrieval-augmented generation. He received the ACM SIGKDD Disertation Award Honorable Mention, Best Paper Award in the 2020 Smoky Mountain Data Challenge Competition by ORNL and GL-Frontiers Workshop at Neurips'23, Best Doctoral Forum Poster Runner-ups at SDM'24, and Outstanding Reviewer at ECML-PKDD.

Xiusi Chen is a Postdoctoral Research Fellow at the University of Illinois Urbana-Champaign, working with Prof. Heng Ji. He received his Ph.D. in Computer Science at the University of California, Los Angeles, advised by Prof. Wei Wang. Xiusi's research focuses on enhancing LLM reasoning, alignment, and decision-making. Xiusi has been awarded the SDM Best Poster Award Honorable Mention. His research has generated over 50 publications in top-tier venues in the fields of data mining, natural language processing, machine learning and information retrieval. He has organized workshops at top venues such as KDD, and has been invited as a keynote/tutorial speaker at NAACL and KDD workshops.

Qi Zhu is an Applied Scientist at AWS, where he focuses on advancing large language model techniques by integrating structured knowledge to multiple applications. He received his Ph.D. from the University of Illinois Urbana-Champaign, advised by Prof. Jiawei Han. He has published over 25 papers in top machine learning and data mining conferences and journals. His research received the 2018 WWW Best Poster Honorable Mention and the 2020 Amazon Machine Learning Research Award.

Yuan Luo is Chief AI Officer at Clinical and Professor at Feinberg School of Medicine in Northwestern University. Globally recognized for his leadership in healthcare AI, Dr. Luo has been elected as Fellow of the American Institute for Medical and Biological Engineering (AIMBE), Fellow of the American College of Medical Informatics (ACMI), Fellow of the American Medical Informatics Association (AMIA), and Fellow of the International Academy of Health Sciences Informatics (IAHSI). His research has been featured in leading journals, including JAMA, Nature Medicine, and Nature Biotechnology and top AI conferences, including NeurIPS, AAAI, ICLR, IJCAI etc.

B Confirmed Program Committee

Ziqing Wang, Ruiyao Xu, Kexin Zhang, Chongyang Gao, Yongjia Lei, Zhisheng Qi, Utkarsh Sahu, Bo Ni, Shanyong Wang, Peixuan Han, Ojas Nimase, Cheng Cheng, Tiankai Yang, Shawn Li, Yuehan Qin, Jiate Li, Hangxiao Zhu, Yuyang Bai, Zhuofeng Li, Haixiang Tang, Fang Guo, Hui Chen, Rongcan Pei, Yikuan Li

C Paper Submission

The workshop will accept submissions through two channels: ACL Rolling Review (ARR) and its own review process via OpenReview.