

2025 IEEE 41st International Conference on Data Engineering Workshops (ICDEW 2025)

**Hong Kong
19-23 May 2025**



**IEEE Catalog Number: CFP2545A-POD
ISBN: 979-8-3315-9960-7**

**Copyright © 2025 by the Institute of Electrical and Electronics Engineers, Inc.
All Rights Reserved**

Copyright and Reprint Permissions: Abstracting is permitted with credit to the source. Libraries are permitted to photocopy beyond the limit of U.S. copyright law for private use of patrons those articles in this volume that carry a code at the bottom of the first page, provided the per-copy fee indicated in the code is paid through Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923.

For other copying, reprint or republication permission, write to IEEE Copyrights Manager, IEEE Service Center, 445 Hoes Lane, Piscataway, NJ 08854. All rights reserved.

****** This is a print representation of what appears in the IEEE Digital Library. Some format issues inherent in the e-media version may also appear in this print version.***

IEEE Catalog Number:	CFP2545A-POD
ISBN (Print-On-Demand):	979-8-3315-9960-7
ISBN (Online):	979-8-3315-9959-1
ISSN:	1943-2895

Additional Copies of This Publication Are Available From:

Curran Associates, Inc
57 Morehouse Lane
Red Hook, NY 12571 USA
Phone: (845) 758-0400
Fax: (845) 758-2633
E-mail: curran@proceedings.com
Web: www.proceedings.com

CURRAN ASSOCIATES INC.
proceedings
.com

2025 IEEE 41st International Conference on Data Engineering Workshops (ICDEW) **ICDEW 2025**

Table of Contents

1st Workshop on Data-AI Systems (DAIS'25)

SEED: Enhancing Text-to-SQL Performance and Practical Usability Through Automatic Evidence Generation	1
<i>Janghyeon Yun (Seoul National University, Korea) and Sang-goo Lee (Seoul National University, Korea; IntelliSys, Korea)</i>	
Fine-Tuning Large Language Models for Entity Matching	9
<i>Aaron Steiner (University of Mannheim, Germany), Ralph Peeters (University of Mannheim, Germany), and Christian Bizer (University of Mannheim, Germany)</i>	
Orchestrating Agents and Data for Enterprise: A Blueprint Architecture for Compound AI	18
<i>Eser Kandogan (Megagon Labs, USA), Nikita Bhutani (Megagon Labs, USA), Dan Zhang (Megagon Labs, USA), Rafael Li Chen (Megagon Labs, USA), Sairam Gurajada (Megagon Labs, USA), and Estevam Hruschka (Megagon Labs, USA)</i>	
Data Cleaning using Large Language Models	28
<i>Shuo Zhang (Columbia University, USA), Zezhou Huang (Columbia University, USA), and Eugene Wu (Columbia University, USA)</i>	
SLPerf: A Research Library and Benchmark Framework for Split Learning	33
<i>Zhanyi Hu (East China Normal University, China), Tianchen Zhou (East China Normal University, China), Bingzhe Wu (Tencent AI Lab, China), Cen Chen (East China Normal University, China), and Yanhao Wang (East China Normal University, China)</i>	
Agentic Workflows for Extraction of Access Control Matrices from Policy Documents	37
<i>Pranav Subramaniam (University of Chicago) and Sanjay Krishnan (University of Chicago)</i>	
Towards Regaining Control over Messy Machine Learning Pipelines	42
<i>Stefan Grafberger (BIFOLD & TU Berlin), Hao Chen (BIFOLD & TU Berlin), Olga Ovcharenko (BIFOLD & TU Berlin), and Sebastian Schelter (BIFOLD & TU Berlin)</i>	

6th International Workshop on Data-Driven Smart Cities (DASC'25)

All About Digitraffic: Understanding Nationwide Open Road Data in Finland	47
<i>Henna Tammia (University of Oulu, Finland), Benjamin Kämä (University of Oulu, Finland), and Ella Peltonen (University of Oulu, Finland)</i>	
Object Detection in Real-World Smart City Applications: A Case of Truck Detection in California Highways	55
<i>Bhavyesh Sajja (University of Southern California, USA), Wenting Qiu (University of Southern California, USA), Jooyoung Yoo (University of Southern California, USA), and Seon Ho Kim (University of Southern California, USA)</i>	
Developing a Smart Electric Vehicle Strategy: From Data to Decisions	63
<i>Laura Greenstreet (Cornell University), Eugenie Y. Lai (MIT), Kevin Lin (University of British Columbia), G. Alexi Rodriguez-Arelis (University of British Columbia), and Raymond Ng (University of British Columbia)</i>	
A Predictive Parking Mobile Application Integrating Historical Usage and Precipitation Data	71
<i>Huijia Wang (Northeastern University, Canada), Ke Wang (Northeastern University, Canada), Tianhao Zhang (Northeastern University, Canada), Yuming Sun (Northeastern University, Canada), Sarita Singh (Northeastern University, USA), and Mario A. Nascimento (Northeastern University, Canada)</i>	

Joint International Workshop on Big Data Management on Emerging Hardware and Data Management on Virtualized Active Systems (HardBD & Active'25)

An Experimental Study of GPU-Based Graph ANN Search Algorithms	79
<i>Yinzuo Jiang (University of Chinese Academy of Sciences) and Shimin Chen (University of Chinese Academy of Sciences)</i>	
On-Chip Acceleration for Log-Structured GDBMS	86
<i>Alexander Baumstark (TU Ilmenau, Germany) and Kai-Uwe Sattler (TU Ilmenau, Germany)</i>	

1st International Workshop on Coupling of Large Language Models with Vector Data Management (LLM+Vector Data)

LLM + Vector Data: Coupling of Large Language Models with Vector Data Management for Enhancing Data Science	93
<i>Arijit Khan (Aalborg University, Denmark), Yuxiang Wang (Hangzhou Dianzi University, China), Weixi Zhang (Huawei Technologies Co., Ltd., China), Yao Tian (Hong Kong University of Science and Technology, China), and M. Tamer Özsu (University of Waterloo, Canada)</i>	

Enhancing Large Language Models with Pseudo- and Multisource-Knowledge Graphs for Open-ended Question Answering	97
<i>Jiaxiang Liu (Chinese Academy of Sciences, China; University of Chinese Academy of Sciences, China), Tong Zhou (Chinese Academy of Sciences, China), Yubo Chen (Chinese Academy of Sciences, China; University of Chinese Academy of Sciences, China), Jun Zhao (Chinese Academy of Sciences, China; University of Chinese Academy of Sciences, China), and Kang Liu (Chinese Academy of Sciences, China; University of Chinese Academy of Sciences, China; Shanghai Artificial Intelligence Laboratory, China)</i>	
LLM-MS: A Multi-Model LLM Search Engine	107
<i>Konstantin Krasovitskiy (University of Cyprus, Cyprus), Stelios Christou (University of Cyprus, Cyprus), and Demetrios Zeinalipour-Yazti (University of Cyprus, Cyprus)</i>	
Campus: Making Cluster-Based Vector Index Scalable with Transactions	115
<i>Masaru Uchida (Keio University) and Hideyuki Kawashima (Keio University)</i>	

2nd International Workshop on Multivariate Time Series Analytics (MulTiSA'25)

Variable Spatiotemporal Framework for Multivariate Time Series Prediction	122
<i>Chen Xu (Beijing University of Posts and Telecommunications), Qiang Wang (Beijing University of Posts and Telecommunications), Yiyang Wu (DiDi Global Inc.), and Lianxing Li (DiDi Global Inc.)</i>	
Impute4TSC: Evaluating Missing Value Imputation Methods for Time Series Classification	127
<i>Xiaou Ding (Harbin Institute of Technology), Cong Peng (Harbin Institute of Technology), Muyun Zhou (Harbin Institute of Technology), Hongzhi Wang (Harbin Institute of Technology), Zhongyi Pei (Tsinghua University), Chen Wang (Tsinghua University), and Jianmin Wang (Tsinghua University)</i>	
User-Friendly Foundation Model Adapters for Multivariate Time Series Classification	136
<i>Romain Ilbert (Huawei Noah's Ark Lab, France; Université Paris-Cité, France), Vasilii Feofanov (Huawei Noah's Ark Lab, France), Malik Tiomoko (Huawei Noah's Ark Lab, France), Ievgen Redko (Huawei Noah's Ark Lab, France), and Themis Palpanas (Université Paris-Cité, France)</i>	
ChronoTab: Forecasting Multivariate Time Series with Tabular LLMs	145
<i>Alexandros Zeakis (National Kapodistrian University of Athens, Greece), Giorgos Chatzigeorgakidis (Athena Research Center, Greece), Konstantinos Lentzos (Athena Research Center, Greece), and Dimitrios Skoutas (Athena Research Center, Greece)</i>	
MULISSE: Variable-Length Similarity Search for Multivariate Time Series	154
<i>Balázs Pelok (Eindhoven University of Technology, the Netherlands) and Jens E. d'Hondt (Eindhoven University of Technology, the Netherlands)</i>	

1st International Workshop on Cognitive and Mental Health Disorder Detection on Social Media: Benchmarking, Data Management, and Data Mining (CMHSM'25)

A Survey of Large Language Models in Mental Health Disorder Detection on Social Media	164
<i>Zhuohan Ge (The Hong Kong Polytechnic University), Nicole Hu (The Chinese University of Hong Kong), Darian Li (The Hong Kong Polytechnic University), Yubo Wang (Hong Kong University of Science and Technology), Shihao Qi (The Hong Kong Polytechnic University), Yuming Xu (The Hong Kong Polytechnic University), Han Shi (Hong Kong University of Science and Technology), and Jason Zhang (The Hong Kong Polytechnic University)</i>	
Deep Learning in Depression Detection: A Comprehensive Survey and Critical Analysis	177
<i>Yuliang Zhang (The Hong Kong Polytechnic University), Jieshun You (The Hong Kong Polytechnic University), Chenghe Yang (The Hong Kong Polytechnic University), Zhumeng Wang (The Hong Kong Polytechnic University), Xixi Qiu (The Hong Kong Polytechnic University), and Yiyu Chen (The Hong Kong Polytechnic University)</i>	
RSD-15K: A Large-Scale User-Level Annotated Dataset for Suicide Risk Detection on Social Media	190
<i>Shouwen Zheng (The Hong Kong Polytechnic University, Hong Kong), Yingzhi Tao (Anhui University, China), and Taiqi Zhou (The Hong Kong Polytechnic University, Hong Kong)</i>	
Advancing Mental Health Research with Graph Neural Networks: A Comprehensive Survey	197
<i>Wei Ma (The Hong Kong Polytechnic University), Zihan Su (The Hong Kong Polytechnic University), Qianwei Chen (The Hong Kong Polytechnic University), Hanshu Zhai (The Hong Kong Polytechnic University), Juanyuan Jiang (The Hong Kong Polytechnic University), and Han Shi (The Hong Kong University of Science and Technology)</i>	
Holistix: A Dataset for Holistic Wellness Dimensions Analysis in Mental Health Narratives	211
<i>Heba Shakeel (Jamia Millia Islamia, India), Tanvir Ahmad (Jamia Millia Islamia, India), and Chandni Saxena (The Chinese University of Hong Kong, China)</i>	

Data Engineering Meets Large Language Models: Challenges and Opportunities (DExLLM'25)

GReaTER: Generate Realistic Tabular Data After Data Enhancement and Reduction	218
<i>Tung Sum Thomas Kwok (University of California, USA), Chi-Hua Wang (University of California, USA), and Guang Cheng (University of California, USA)</i>	
Hallucination Detection with Small Language Models	230
<i>Ming Cheung (The Lane Crawford Joyce Group, China)</i>	
Risk-Aware Automatic Text Summarization with Pre-Identification of Risk Categories and Emphasis	239
<i>Hailin Huang (South China University of Technology; MOE of China), Hongfei Liu (South China University of Technology; MOE of China), Xin Wu (South China University of Technology; MOE of China), and Yi Cai (MOE of China; South China University of Technology)</i>	

RBRTI: Retrieval-Based Risk Type Identification from Financial News	246
<i>Songwen Gong (South China University of Technology; MOE of China), Hongfei Liu (South China University of Technology; MOE of China), Xin Wu (South China University of Technology; MOE of China), and Yi Cai (MOE of China; South China University of Technology)</i>	
Structured Retrieval-Augmented Generation for Multi-Entity Question Answering over Heterogeneous Sources	253
<i>Teng Lin (HKUST(GZ), China)</i>	
Retrieval-Augmented Classification for Financial User Profiling: A Lightweight Approach	259
<i>Sile Guo (South China University of Technology; Ministry of Education), Hongfei Liu (South China University of Technology; Ministry of Education), Xin Wu (South China University of Technology; Ministry of Education), and Yi Cai (South China University of Technology; Ministry of Education)</i>	
ZiGong 1.0: A Large Language Model for Financial Credit	266
<i>Yu Lei (Didi International Business, China), Zixuan Wang (Didi International Business, China), Chu Liu (Didi International Business, China), and Tongyao Wang (Didi International Business, China)</i>	

8th International Workshop on Data Engineering Meets Intelligent Food and Cooking Recipes (DECOR'25)

RICGraph: A Recipe-Ingredient-Compound Graph for Estimating the Nutritional Value of Recipes	271
<i>Naoki Yoshimaru (Doshisha University, Japan), Kazuma Kusu (Doshisha University, Japan), Yusuke Kimura (Doshisha University, Japan), and Kenji Hatano (Doshisha University, Japan)</i>	
A Multilingual Ontology and Knowledge Graph for Recipes	279
<i>Mansi Goel (Indraprastha Institute of Information Technology Delhi (IIIT-Delhi), Inde) and Frederic Andres (National Institute of Informatics, Japan)</i>	
IEML—A New Frontier for Semantic Computing and Collective Intelligence in the Cooking Domain	286
<i>Mansi Goel (IIIT-Delhi, India), Frederic Andres (National Institute of Informatics, Japan), and Pierre Levy (University of Montreal, Canada)</i>	

1st International Workshop on Quantum Data and Machine Learning (QDML'25)

Optimizing Join Orders via Constrained Quadratic Models (Abstract)	292
<i>Hanwen Liu (University of Southern California, USA), Pranshi Saxena (University of Southern California, USA), Federico Spedalieri (University of Southern California, USA), and Ibrahim Sabek (University of Southern California, USA)</i>	

Data-Centric Approach of Macroscopic Physical System with NISQ Quantum Computers	293
<i>Junyong Lee (Yonsei University, South Korea), Jeihee Cho (Yonsei University, South Korea), Hyeonseong Jung (Yonsei University, South Korea), Euimin Lee (Yonsei University, South Korea), Sangmin Lee (Yonsei University College of Medicine, South Korea), Yunah Choi (Yonsei University, South Korea), and Shiho Kim (Yonsei University, South Korea)</i>	
Neural Quantum Embeddings with Multiple Control Variables	298
<i>SeungYeop Baik (Yonsei University, Republic of Korea) and Yo-Sub Han (Yonsei University, Republic of Korea)</i>	
Author Index	303