

DocInsights 2026

Workshop on Document Intelligence and Understanding

Proceedings of the Workshop

October 24, 2026

©2026 Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-477-4

Preface

Welcome to the first Workshop on Document Intelligence and Understanding (DocInsights 2026), held in Budapest, Hungary, on October 24, 2026, co-located with EMNLP 2026.

Documents such as reports, forms, financial statements, scientific papers, and manuals are among the richest sources of information in science, industry, and society, combining text with tables, figures, and complex layouts. DocInsights brings together researchers and practitioners working on document understanding, from parsing and information extraction to multimodal retrieval, question answering, and reasoning over long and visually rich documents, with a focus on reliable, evidence-grounded systems.

This year, the main workshop track received 82 direct submissions (excluding one withdrawn submission) and 18 papers committed through ACL Rolling Review (ARR). We accepted 39 of these 100 papers (39%): 29 appear in these proceedings, and 10 non-archival papers are presented at the workshop only. The accepted papers cover document parsing and OCR robustness, table understanding and extraction, multimodal and long-document retrieval-augmented generation, visual document retrieval, evaluation methodology for document AI, agents for document and table processing, and applications to financial, biomedical, policy, and scientific documents.

The workshop also hosted two shared tasks: DocSem, on evidence-grounded quantitative reasoning over documents, and Dr.DocBench, on expert-level document parsing. We accepted 12 system description papers, 5 of which appear in these proceedings under Shared Task Papers.

We are grateful to the 147 members of our program committee for their careful and timely reviews, and to all the authors and shared task participants. We especially thank our shared task organizers, Jyotika Singh, Karan Dua, Hitesh Patel, Rahul Suresh, Xiaolong Luo, and Alexandra Bezea-Tudor, for designing and running the DocSem and Dr.DocBench challenges.

We hope you enjoy the workshop!

The DocInsights 2026 Organizers

Santosh T.Y.S.S., Amit Agarwal, Vivek Gupta, Yova Kementchedjheva, Aparna Garimella, and Yufang Hou

Organizing Committee

Organizers

Santosh T.Y.S.S., Amazon, USA
Amit Agarwal, Oracle, USA
Vivek Gupta, Arizona State University, USA
Yova Kementchedjhieva, MBZUAI, UAE
Aparna Garimella, Adobe Inc, India
Yufang Hou, IT:U Austria / IBM Research Europe

Shared Task Organizers

Jyotika Singh, Oracle AI
Karan Dua, Oracle AI
Hitesh Patel, Oracle AI
Rahul Suresh, Abaka AI
Xiaolong Luo, Harvard University
Alexandra Bezea-Tudor, Abaka AI

Program Committee

Program Committee

Amirhossein Abaskohi, University of British Columbia
Mahmud Elahi Akhter, Queen Mary, University of London
Tejas Anvekar, Arizona State University
Longju Bai, University of Michigan - Ann Arbor
George Arthur Baker, University of Chicago
Bal Krishna Bal, Kathmandu University
Nikita Bhutani, Megagon Labs, Inc
Zahra Bokaei, University of Edinburgh
Shubin Cai, Shenzhen University
Xinyu Cai, Shanghai Artificial Intelligence Laboratory
Lele Cao, Microsoft
Yixuan Cao, Institute of Computing Technology, Chinese Academy of Sciences
Abhijit Chakraborty, Arizona State University and MongoDB
Jiayi Chen, Fudan University
Ming-Bin Chen, University of Melbourne
Xinchi Chen, Fudan University
Zesen Cheng, Peking University
Purva Chiniya, Amazon
Jeonghwan Choi, Korea Advanced Institute of Science & Technology
Jihun Choi, Knowledge Work Inc.
Anna Corazza, Università Federico II di Napoli, Italy
Xiang Dai, CSIRO
Preetam Prabhu Srikar Dammu, University of Washington
Yufan Dang, Tsinghua University
Pavel Denisov, Fraunhofer IAIS
Hang Ding, Shanghai Jiaotong University
Wentao Ding, Beijing Institute for General Artificial Intelligence
Hervé Déjean, Naver Labs Europe
Xiran Fan, VISA
Ian T. Foster, University of Chicago and Argonne National Laboratory
Sheela Ramprakash Goenka, Google Deepmind
Boyu Gou, Ohio State University, Columbus
Ray Groshan, University of Maryland, Baltimore County
Sonam Gupta, International Business Machines
Masoud Hashemi, ServiceNow Inc
Hilaf Hasson, Cohesity
Qijia He, University of Washington
Pingjun Hong, Universität Vienna
Arya Honraopatil, University of Maryland, Baltimore County
Jun Hou, Virginia Polytechnic Institute and State University
Bodun Hu, University of Texas at Austin
Wei-Chieh Huang, University of Illinois at Chicago
Jena D. Hwang, Allen Institute for Artificial Intelligence
Pooja Jhunjhunwala, Google
Yuelyu Ji, University of Pittsburgh
Ziyou Jiang, Institute of Software Chinese Academy of Sciences

Ruochun Jin, National University of Defense Technology
 Gyuwan Kim, University of California, Santa Barbara
 Taehyeong Kim, Seoul National University
 Dongyuan Li, The University of Tokyo and Institute of Science Tokyo
 Rongchun Li, National University of Defense Technology
 Ruosen Li, University of Texas at Dallas
 Ryan Li, Tesla AI
 Yantao Li, nanjing university
 Yinghao Li, Shanxi University
 Zhuoqun Li, University of the Chinese Academy of Sciences
 Ruixue Lian, Amazon
 I-Bin Liao, Hon Hai Technology Group (Foxconn®)
 Xinting Liao, University of British Columbia and Vector Institute
 Shanshan Liu, National Institute for Materials Science and University of Tsukuba, Tsukuba University
 Yunfei Long, Queen Mary, University of London
 Shangyu Lou, San Diego State University and University of California, Santa Barbara
 Jianfeng Lu, Nanjing university of Science & Technology
 Guoqing Luo, University of Alberta
 Haodi Ma, University of Florida
 Khyati Mahajan, ServiceNow Inc
 Harsh Maheshwari, Amazon
 Ziqing Man, UCLA Computer Science Department, University of California, Los Angeles
 Qianren Mao, Zhongguancun Laboratory, Beijing, P.R.China. and Beihang University
 Chunyu Miao, University of Illinois at Chicago
 Zhongtao Miao, The University of Tokyo
 Ishani Mondal, University of Maryland, College Park
 Noriki Nishida, AIST, National Institute of Advanced Industrial Science and Technology
 Yusuke Oda, National Institute of Informatics and Nara Institute of Science and Technology
 Hamidah Oderinwale, McGill University
 Naoki Otani, Megagon Labs
 Mingyu Ouyang, National University of Singapore
 Matteo Paganelli, University of Modena and Reggio Emilia
 Jeongsik Park, GE HealthCare
 Dhaval C Patel, International Business Machines
 Xinyu Pi, University of California, San Diego
 Xin Quan, University of Manchester
 Saman Rahbar, Dialpad AI, Nurty Technologies Inc, University of Calgary and University of British Columbia
 Sourajit Saha, University of Maryland, Baltimore County
 Tobias Schimanski, University of Zurich
 Nisarg Shah, Uber
 Arun Sharma, University of Minnesota - Twin Cities
 Neelesh Kumar Shukla, Oracle
 Vaibhavi Singh, New York University
 Sebastian Stüker, Zoom Video Communications
 Xiangguo Sun, Southeast University
 Zexu Sun, Baidu
 Zhouhao Sun, Harbin Institute of Technology
 Jing Tang, Huazhong University of Science and Technology
 Wei Tao, LIGHTSPEED

Jinlong Tian, National University of Defense Technology
 Jannis Vamvas, University of Zurich
 Benlu Wang, Department of Computer Science, University of Massachusetts at Amherst
 Chenan Wang, College of William and Mary
 Danqing Wang, CMU, Carnegie Mellon University
 Haocheng Wang, The Hong Kong University of Science and Technology and ETHZ - ETH Zurich
 Haoran Wang, Emory University
 Juanyan Wang, Netradyne
 Shanyong Wang, Xiaohongshu
 Wei Wang, Nankai University
 Xiaobo Wang, Beijing Institute for General Artificial Intelligence and University of Science and Technology of China
 Yichen Wang, Georgia Institute of Technology
 Zixuan Wang, Zhejiang University and Baidu
 Xing Wei, Xi'an Jiaotong University
 Tianhao Wu, Tianjin University
 Yinjun Wu, Peking University
 Yuankai Wu, Sichuan University
 Zengqing Wu, Osaka University
 Zhiheng Wu, Baidu
 Meng Xiao, National University of Singapore and Chinese Academy of Sciences
 Ke Xu, University of Liverpool
 Pei Xu, Institute of automation, Chinese academy of science, Chinese Academy of Sciences
 Zhen Xu, PCG Tencent
 Jiangpeng Yan, E Fund Management Co., Ltd.
 Ming Yan, Xinjiang University
 Jinfa Yang, Fuzhou University
 Tianyu Yang, Georg-August Universität Göttingen
 Xin Yang, Southwestern University of Finance and Economics
 Zhiyu Yao, Tsinghua University
 Yangfan Ye, Harbin Institute of Technology
 Ryota Yoshihashi, Institute of Science Tokyo and LY Corporation
 Hang Yu, Shanghai University
 Ruihong Zeng, Mohamed bin Zayed University of Artificial Intelligence
 Chi Zhang, Beijing Institute of Technology
 Han Zhang, Renmin University of China
 Jinghui Zhang, Mohamed bin Zayed University of Artificial Intelligence
 Liang Zhang, SUN YAT-SEN UNIVERSITY
 Pingping Zhang, Dalian University of Technology
 Shuyi Zhang, Wuhan University
 Xitong Zhang, Qualcomm Inc, Qualcomm
 Yunqi Zhang, Zhongguancun Laboratory
 Zhong Zhang, University of Electronic Science and Technology of China
 Ziqi Zhang, Institute of automation, Chinese academy of science, Chinese Academy of Sciences
 Kaiyan Zhao, The University of Tokyo
 Yang Zhao, Tsinghua University, Tsinghua University
 Yi Zhao, Harbin Institute of Technology
 Zhenjie Zhao, Nankai University
 Zijian Zhao, The Hong Kong University of Science and Technology
 Juepeng Zheng, SUN YAT-SEN UNIVERSITY
 Guochen Zhou, City University of Hong Kong

Jiawei Zhou, Fudan University
Zirui Zhuang, Beijing University of Posts and Telecommunications

Table of Contents

<i>A Human-Annotated Reference Corpus for Medication Information Extraction from MIMIC-IV Discharge Summaries</i>	
Provia Kadusabe, Abhishek Kaushik and Fiona Lawless	1
<i>The Right Information Extraction Pipeline Depends on the Document: Accuracy-Energy Trade-offs for Small, Local Models</i>	
Christoph Marjan Markus Walser, Mauricio Fadel Argerich and Jonathan Fürst	12
<i>Splitting Documents at Lower Cost: Multi-Split Boundary Decisions for LLM-Based Page Stream Segmentation</i>	
Nikhil Reddy Pottanigari, Sepideh Kharaghani, Saverio Vadacchino, Alejandro Posada and Ying Zhang	30
<i>Text, Pixels, or Both? Evaluating Input Representations for Multimodal Document QA</i>	
Nikhil Reddy Pottanigari, Sepideh Kharaghani, Saverio Vadacchino, Alejandro Posada, Kurt MacDonald and Ying Zhang	46
<i>AdaptiveDocRAG: Efficient Document Understanding with Multimodal Large Language Models</i>	
Surya Selvam and Anand Raghunathan	63
<i>Most of a Document-VQA Score Is Answer Formatting</i>	
Srinivas Harish	73
<i>Graph-Guided Passage Retrieval for Author-Centric Structured Feedback</i>	
Maitreya Prafulla Chitale, Ketaki Mangesh Shetye, Harshit Gupta, Manav Chaudhary, Manish Shrivastava and Vasudeva Varma	81
<i>GraphER: An Efficient Graph-Based Enrichment and Reranking Method for Retrieval-Augmented Generation</i>	
Ruizhong Miao, Yuying Wang, Rongguang Wang, Chenyang Li, Tao Sheng, Sujith Ravi and Dan Roth	100
<i>LongDA: Benchmarking LLM Agents for Long-Document Data Analysis</i>	
Yiyang Li, Zheyuan Zhang, Tianyi Ma, Zehong Wang, Keerthiram Murugesan, Chuxu Zhang and Yanfang Ye	112
<i>Reflect and Retrieve: Multimodal Agentic Retrieval with Reflection and Query Agents for Document Search</i>	
Tryambak Gangopadhyay, Md Sirajus Salekin, Yiyue Qian, Nkechinyere Nneka Agu, Shinan Zhang and Rinne Han	138
<i>LLMs and Reasoning in Health Insurance Policy Documents: A Dataset and Argument Graph Extraction Study</i>	
Govind Arun, Sidharth N Krishna, Susie Xi Rao and Ajo Babu George	155
<i>NutriChat: Evaluating Safety-Aware Retrieval-Augmented Generation for Nutrition Textbook Question Answering</i>	
Hadis Ghafouri	163
<i>Measuring Artifact Reliance in Machine-Generated Text Detectors</i>	
Anastasia Voznyuk and Andrey Grabovoy	174

<i>Graph Retrieval in Multimodal RAG for Biomedical Documents: A Pharmacokinetics and Pharmacodynamics Case Study</i>	
Primoz Kocbek, Katarina Meštrović, Nguyen Thu Huong, Ales Fajmut and Gregor Stiglic . . .	186
<i>Q-STR: Diagnosing Query-Driven Text-to-Table Reasoning in Cricket Narratives</i>	
Ritam Upadhyay, Naman Ahuja, Rishabh Baral, Aparna Garimella and Vivek Gupta	196
<i>Ontology-Guided Procedural Knowledge Extraction and Validation Case Study on the Lockout-Tagout Documents</i>	
Cennet Oguz, Noon Pokaratsiri Goldstein, Jörg Steffen, Reetam Biswas and Günter Neumann	217
<i>Perception, Layout, and Validation: Calibrated Confidence for Reliable Straight-Through Processing of Financial Documents</i>	
Yichao Jin, Yushuo Wang, Han Yuxuan, Ching Yee Sonia Kwan, Weiyang Song, Kent Chiu Jin Chun, Chong Hwee Wong, Tiong Kiat Wong, Kenneth Zhu Ke and Jingyuan Zhao	232
<i>Tabstructor: Reconstructing Latent Borders in Borderless and Partially Ruled Tables</i>	
Alvin Sarraga Alon, Daniele De Rosa, Johannes Friedl and Riza Batista-Navarro	244
<i>Silent Structured-Content Corruption in Document Parsing: A Span-Typed Diagnostic and a Released Cross-Parser Resource</i>	
Ruiyang Zhang	273
<i>Sorry Robot, Happy Human: Vision-Language Models Read Only One of Two Legible Typographic Layers</i>	
Mert İncidelen, Yamen Kashkash, Asya Berker and Murat Aydogan	282
<i>What Should Occupy the Context Budget? Latent Compression versus Budget-Matched Truncation in Document QA</i>	
Artem Cherepanov and Albina Burlova	287
<i>RAG Components Are Not Independently Rankable</i>	
Vishnu Prasad Vijaya Kumar, Iaroslav Neverov, Pavel Chizhov, Anton Changalidis, Vaiva Kazanavičiute, Pierre-Carl Langlais, Anastasia Stasenko and Ivan P. Yamshchikov	303
<i>LaTeXTrans Structure-Aware Translation of Complex LaTeX Documents</i>	
Ziming Zhu, Chenglong Wang, Haosong Xu, Shunjie Xing, Yifu Huo, Fengning Tian, Quan Du, Di Yang, Chunliang Zhang, Tong Xiao and JingBo Zhu	324
<i>Did You CoSeCh?: Faithfulness- and Coherence-Aware Auditing of Content Selection in Poster Generation</i>	
Aaditya Deshpande, Savita Bhat and Shirish Karande	344
<i>Look Where It Matters: High-Resolution Crops Retrieval for Efficient VLMs</i>	
Nimrod Shabtay, Moshe Kimhi, Sivan Harary, Artem Spector, Ehud Rivlin, Chaim Baskin, Raja Giryes and Eli Schwartz	357
<i>DistilVDR: A Compact End-to-End Visual Document Retriever via Dual-Student Distillation</i>	
Zhuchenyang Liu, Ziyi Wang, Yao Zhang and Yu Xiao	385
<i>Comparative Study of AI Agent Architecture Efficiency in Table Processing Tasks</i>	
Timur Gabbasov, Aleksandr Sandalov, Aleksandr Maksimov, Maria Khodorchenko and Nikolay Alekseevich Butakov	400
<i>MaDOVE: a Dataset for Multilingual Document VQA Evaluation</i>	
Elio Musacchio, Lucia Siciliani, Pierpaolo Basile, Giovanni Semeraro, Mark Federico Zampedroni and Asia Beatrice Uboldi	409

Generating Edit-Inducing Questions for AI Research Manuscripts

Sebastian Antony Joseph, Zichao Wang, Jennifer Healey, Alexa Siu, Junyi Jessy Li and Ani Nenkova 428

Shared Task Papers

When the Evidence Is Right, the Math Is Easy: Understanding the Challenges of Document-Grounded Reasoning

Aftab Hossain Mikat and Tamjid Hasan Fahim 440

Acessilia at Dr.DocBench 2026: Agentic Document Parsing through Canonical Tree Adjudication

Marcelo Akira Inuzuka, Marcos Paulo C M Queiroz, Jhonata Fernandes Cordeiro and Eduardo Fraga Pereira 448

Evidence First, Arithmetic Second: A System Report and Failure Analysis for DOCSEM

Divya Godara and Sachin Gupta 460

Retrieve, Reason, Verify: A Modular Multi-Model System for DocSem

Pranavi Tadivalasa 465

A Cost-Efficient Vision-Text Cascade for Document QA: What Resolution Buys and What It Doesn't

Urjita Bedekar 477