T4: Multimodal Large Language Models for Human–AI Interaction: Foundations, Agents, and Inclusive Applications

Shafiq Joty, Enamul Hoque, Ahmed Masry,
Spandana Gella, Samira Ebrahimi Kahou
https://mllm4haii.github.io/
Cutting-edge
Salle Le Riad
Sunday, March 29, 2026 - from 09:00 to 12:30

Multimodal large language models (MLLMs) are redefining how humans communicate and collaborate with machines. They extend the capabilities of text-based LLMs to perceive, reason, and act across text, images, charts, forms, and graphical user interfaces (GUIs). These models are now capable of answering questions about charts, summarizing infographics, operating software through natural language, and supporting multilingual and accessible visualization.

This tutorial offers a concise, three-hour introduction to the foundations, agentic capabilities, and inclusive applications of MLLMs, with a focus on visually grounded and interactive language tasks. We will cover core architectural designs (encoders, connectors, fusion and decoding mechanisms), multimodal alignment and learning strategies, and reasoning techniques for structured visuals such as charts, forms, and infographics. The tutorial then examines multimodal and conversational agents that perform dialogue-driven reasoning and co-creative analysis in graphical user interfaces. We conclude with discussions on accessibility, multilingual communication, responsible deployment, and future challenges in building human-centered multimodal AI.

Shafiq Joty, Salesforce Research, USA
email: sjoty@salesforce.com
website: https://raihanjoty.github.io/
Bio. Shafiq Joty is a Research Director at Salesforce Research, and is also an Associate Professor (on leave) at NTU, Singapore. His work has primarily focused on developing language analysis tools and NLP applications. A significant part of his current research focuses on multilingual (machine translation, cross-lingual

transfer), multimodal (visual-language learning, NLP+Vis, Code+NLP) NLP, interpretability and robustness of NLP models. His research contributed to 17 patents and more than 110 papers in top-tier NLP and ML conferences and journals including ACL, EMNLP, NAACL, NeurIPS, ICML, ICLR, CVPR, ECCV, ICCV, CL and JAIR. Shafiq served (or will serve) as a PC chair of SIGDIAL'23, an S/AC for ICLR-23, ACL'22, EMNLP'21, ACL'19-21, EMNLP'19, NAACL'21 and EACL'21 and an AE for ACL-RR. He gave tutorials at IEEE Vis'22, ACL'19, ICDM'18 and COLING'18, and taught deep learning for NLP,[1] a graduate-level NLP course, and an undergraduate NLP course at NTU.

Enamul Hoque, York University, Canada
email: enamulh@yorku.ca
website: https://www.yorku.ca/enamulh/
Bio. Enamul Hoque is an Associate Professor at York University where he directs the Intelligent Visualization Lab. Previously, he was a postdoctoral fellow in Computer Science at Stanford University. He received the Ph.D. degree in Computer Science from the University of British Columbia. His research focuses on combining information visualization and human-computer interaction with natural language processing to address the challenges of the information overload problem. Recently, he has worked on developing natural language interfaces for visualizations, automatic chart question answering, chart retrieval and chart summarization. He has also worked on developing visual text analytics to support the user's task of exploring and analyzing conversations. Since his research is uniquely positioned at the intersection of information visualization, NLP, and HCI, he publishes at the major venues in each of these areas such as IEEE Vis, ACL, EMNLP, CHI, and UIST. He serves as an Area Chair for the ACL Rolling Review (2021-) and as a program committee member (2018-) for the IEEE Vis.

Ahmed Masry, York University, Canada
email: masry20@yorku.ca
website: https://ahmedmasryku.github.io
Bio. Ahmed Masry is a PhD student at York University, Canada, supervised by Professor Enamul Hoque. He previously interned at ServiceNow Research and Mila. His research focuses on developing benchmarks and vision-language models for chart and document understanding, with an emphasis on supervised fine-tuning datasets, reinforcement learning, and improving vision–language alignment in multimodal architectures. Ahmed has led and contributed to popular benchmarks like ChartQA, ChartQAPro and Chart-to-Text, as well as models includ-

[1] https://ntunlpsg.github.io/ce7455_deep-nlp-20/

ing ChartGemma, BigCharts-R1, and AlignVLM. His work has been published in leading NLP and ML venues like ACL, EMNLP, COLM, NeurIPS, ICLR. He also serves as a reviewer for ACL Rolling Review, NeurIPS, and ICLR. Ahmed has received several prestigious Canadian national and provincial awards, including the NSERC CGS-D, Ontario Graduate Scholarship, and Mitacs Accelerate Award. His research was also recognized with the Best Paper Award at the ChartQA Workshop at CVPR 2021.

Spandana Gella, ServiceNow Research, Canada
email: spandana.gella@servicenow.com
website: https://www.servicenow.com/research/author/spandana-gella.html
Bio. Spandana Gella is a Research Scientist at ServiceNow Research. She holds a Ph.D. in Computer Science from the University of Edinburgh. Her research focuses on building robust and safe frontier models and autonomous LLM-agents. In the past, she co-organized multiple workshops co-located with top-tier conferences including Representation Learning for NLP (2018, 2019, 2020), Shortcomings in Vision and Language (2018, 2019), and the Workshop on Multilingual Multimodal Learning (2022).

Samira Ebrahimi Kahou, University of Calgary, Canada
email: samira.ebrahimikahou@ucalgary.ca
website: https://saebrahimi.github.io/
Bio. Samira Ebrahimi Kahou is an Associate Professor at the Electrical and Software Engineering Department at the University of Calgary and an adjunct professor at the School of Computer Science at McGill University. She holds a CIFAR AI Chair. Ebrahimi Kahou's pioneering work in visual reasoning includes the two well-known datasets "Something Something" and "FigureQA". Ebrahimi Kahou and her group currently work on solving fundamental problems in representation learning for decision making, with a broad focus on generalization and efficient learning. Besides this primary focus, she also has expertise in knowledge distillation, climate modeling using deep learning, building large-scale datasets, clinical decision making, and NLP. Her publications appear in leading venues such as NeurIPS, ICLR, ICML, CVPR, and ICCV, and she has served as an Area Chair for NeurIPS, ICCV, and EMNLP.