REALM 2025

The 1st Workshop for Research on Agent Language Models

Proceedings of the Workshop

July 31, 2025

©2025 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 979-8-89176-264-0

Introduction

We are excited to welcome you to REALM 2025, the 1st Workshop for Research on Agent Language Models. The workshop is being hosted by ACL, on July 31, 2025, and will take place in Vienna, Austria.

The REALM Workshop aims to bring together researchers, practitioners, and thought leaders in the ACL community to discuss and align on the current landscape, key challenges, and future directions of LLM agents. The program includes 7 keynote talks, 1 panel discussion, 5 spotlight presentations, and two poster sessions. The workshop will feature a mix of remote and live presentations and discussions, allowing for a diverse and inclusive exchange of ideas.

We received 71 submissions. Every submission was assigned to at least three reviewers. When making our selections for the program, we carefully considered the reviews and the comments made during the discussions among reviewers. The members of the Program Committee did an excellent job in reviewing the submitted papers, and we thank them for their essential role in selecting the accepted papers and helping produce a high quality program for the conference. Our aim has been to create a balanced program that accommodates as many favorably rated papers as possible. We accepted 48 papers: 11 spotlight papers and 37 poster papers. Among the spotlight papers, 5 were selected for short presentations. The acceptance rate for spotlight papers is 15.5% and for poster papers is 52.1%.

A sucessful workshop requires advice, help and enthusiastic participation of many parties, and we have a big 'thank you' to say to all of them. Regarding the program, we thank our seven speakers, Nicolas Chapados (ServiceNow Research), Siva Reddy (McGill University / Mila), Roberta Raileanu (Google DeepMind), Christopher Manning (Stanford University), Tao Yu (University of Hong Kong), Yu Su (Ohio State University), and Daniel Fried (Carnegie Mellon University), for their inspiring talks. We extend special thanks to our ACL Workshop Chair, Terra Blevins and Christophe Gravier, for handling the organization of many workshops in parallel. In addition, we thank the ACL conference for hosting our workshop.

We once again thank our program committee members for committing their time to help us select an excellent technical program. Finally, we thank all the authors who submitted to the workshop and all workshop participants for making REALM 2025 a success and for growing the research areas of LLM Agents with their fine work.

Ehsan Kamalloo, Nicolas Gontier, Xing Han Lu, Shikhar Murty, Alexandre Lacoste and Nouha Dziri
— *Program Co-Chairs*

Organizing Committee

Program Chair

Ehsan Kamalloo, ServiceNow Research
Nicolas Gontier, ServiceNow Research
Xing Han Lu, McGill University and Mila - Quebec AI Institute
Nouha Dziri, Ai2
Shikhar Murty, Stanford University
Alexandre Lacoste, ServiceNow Research

Advisory Board

Hannaneh Hajishirzi, University of Washington / Ai2
Graham Neubig, Carnegie Mellon University / All Hands AI

Program Committee

Program Chairs

Nouha Dziri, Ai2
Nicolas Gontier, ServiceNow Research
Ehsan Kamalloo, ServiceNow Research
Alexandre Lacoste, ServiceNow Research
Xing Han Lù, McGill University and Mila - Quebec AI Institute
Shikhar Murty, Stanford University

Reviewers

Amirhossein Abaskohi, Neelima Agarwal, Tushar Aggarwal, Md Rifat Arefin, Rabiul Awal

Nitin Nikamanth Appiah Balaji, Sahil Bansal, Georges Belanger Albarran, Rishika Bhagwatkar, Aarohi Bhand, Léo Boisvert

Liangliang Chen, Yimeng Chen, Ze Chen, Pranjal A Chitale, Christopher Zhang Cui

Prakhar Dixit, Cong-Thanh Do, Zhengyuan Dong, Alexandre Drouin, Wanyu Du

Youmna Farag, Moghis Fereidouni, Yicheng Fu

Sajjad Ghiasvand, Matej Gradoš, Aylin Ece Gunal

Larry Heck, Lukas Hilgert

Luo Ji, Xiangru Jian

Zong Ke, Arian Khorasani, Byung-Hak Kim, Kemal Kirtac, Sravanth Kodavanti, Mandar Kulkarni

Issam H. Laradji, So Young Lee, Jessica E. Liang, Duoduo Liao, Tobias Lindenbauer, Philip Lippmann, Lajanugen Logeswaran, Peter Lorenz

Shiva Krishna Reddy Malay, Maria Emilia Mazzolenis, Zaiqiao Meng, Aristides Milios, Kyle Moore, Yash More, Kaiyu Mu

Ashish Pandian, Emiliano Penaloza, Leonardo Perelli, Chau Minh Pham, David Guzman Piedrahita, Nearchos Potamitis, Abhay Puri

Santhoshi Ravichandran, Micah Rentschler, Christopher Gordon Richardson, Jesse Roberts, K Roth, Nikolai Rozanov, Jonathan Rystrøm

Matthew Shardlow, Michael Naeim Shehata, Oleh Shliazhko, Noam Slonim, Yixiao Song, Jason Stanley, Hiroaki Sugiyama, Anirudh Sundar

Katherine Thai, Megh Thakkar, Saksham Thakur, Ada Defne Tur

Filippos Ventirozos, Tomás Vergara Browne, Shreyas Verma

Chengcheng Wei, Yunnan Wu

Yinggan XU, R. Patrick Xian, Ying Xie

Zachary Yang, Xunjian Yin, Ori Yoran, Peiyang Yu, Yanxiang Yu, Kamer Ali Yuksel

Chaoyun Zhang, Meiru Zhang, Ruirui Zhang, Tianyu Zhang, Xiao Zhang, Henry Hengyuan Zhao, Xiaofeng Zhu

Keynote Talk

Agentic Full-Stack Benchmarking for Knowledge Work

Nicolas Chapados
ServiceNow Research

Thursday, July 31, 2025 – Time: **09:00 – 09:30** – Room: **Room 1.61-62**

Abstract: In the span of less than one year, AI agents have moved from being a research curiosity to underpinning the largest software platform updates in a generation. Agents promise to streamline substantial portions of knowledge work and progress has been swift, with early-2025 work by METR suggesting a potential doubling of task complexity every seven months. In this talk, we submit this claim to deeper empirical scrutiny: we examine what it takes to build good benchmarks of agentic performance on long-running knowledge work tasks. We review recent contributions, many by the ServiceNow Research team, on task-space dimensions ranging from browser use and multimodal understanding, to data analytics and deep research. We also touch on benchmarks of agentic safety and security, suggesting that their evaluations cannot be decoupled from primary task performance. We arrive at a more nuanced conclusion than earlier results, with major problems in the field yet to be solved by frontier agents.

Bio: Nicolas Chapados is VP of Research at ServiceNow Research. He holds an engineering degree from McGill University and a PhD in Computer Science from University of Montreal, Canada. While still writing his thesis and jointly with his advisor Yoshua Bengio, he co-founded ApSTAT Technologies in 2001, a machine learning technology transfer firm, to apply cutting-edge academic research ideas to areas such as insurance risk evaluation, supply chain planning, business forecasting, national defence, and hedge fund management. From this work, he also co-founded spin-off companies: Imagia, to detect and quantify cancer early with AI analysis of medical images, Element AI, to help organizations plan and implement their AI transformation, and Chapados Couture Capital, a quantitative asset manager. He holds the Chartered Financial Analyst (CFA) designation.

Keynote Talk

The Path to Scalable and Safe Web Agents

Siva Reddy
McGill University / Mila

Thursday, July 31, 2025 – Time: **09:30 – 10:00** – Room: **Room 1.61-62**

Abstract:

Bio: Siva Reddy is an Assistant Professor in the School of Computer Science and Linguistics at McGill University, a Canada CIFAR AI Chair, a core faculty member of Mila Quebec AI Institute, and a research scientist at ServiceNow Research. He co-leads the McGill NLP Group. Previously, he was a postdoctoral researcher at Stanford University and a Google PhD fellow at the University of Edinburgh. His research interests are in representation learning for language with a specific focus on reasoning, grounding, conversational modeling, and safety.

Keynote Talk

Automating Scientific Discovery: How Far Are We?

Roberta Raileanu
Google DeepMind

Thursday, July 31, 2025 – Time: **11:30 – 12:00** – Room: **Room 1.61-62**

Abstract: In this talk, I'll discuss the emergent field of using frontier models such as LLMs for automating scientific discovery and AI research itself. I will first describe the goals of this research area, the various subproblems, proposed approaches, and early work in this space. Despite the hype, flashy news articles, and some recent works with bold claims, I will provide empirical evidence that models still struggle with many aspects of scientific discovery. I argue this is still an open problem and it is unclear whether the current AI paradigm is enough to achieve the long-term ambition of this research agenda. I will then introduce MLGym and MLGym-Bench, a new framework and benchmark for evaluating and developing LLM agents on AI research tasks. This is the first Gym environment for machine learning (ML) tasks, enabling research on reinforcement learning (RL) algorithms for training such agents. MLGym-bench consists of 13 diverse and open-ended AI research tasks from diverse domains such as computer vision, natural language processing, reinforcement learning, and game theory. Solving these tasks requires real-world AI research skills such as generating new ideas and hypotheses, creating and processing data, implementing ML methods, training models, running experiments, analyzing the results, and iterating through this process to improve on a given task. I will demonstrate how MLGym makes it easy to add new tasks, integrate and evaluate models or agents, generate synthetic data at scale, as well as develop new learning algorithms for training agents on AI research tasks. Finally, I will discuss our findings from evaluating frontier LLMs on MLGym-bench, highlighting the limitations of current models at conducting AI Research, as well as avenues for future work.

Bio: Roberta Raileanu is a Senior Staff Research Scientist at Google DeepMind in the Open-Endedness team. She is also an Honorary Lecturer (Adjunct Professor) at UCL, working closely with the UCL-DARK Lab where she advises PhD students. She also co-developed and is currently co-teaching a new course on Open-Endedness and General Intelligence. Previously, she was a Research Scientist at Meta GenAI where she led the AI Scientist team. Her goal was to accelerate AI Research and Development using AI Agents that can generate ideas and hypotheses, implement new methods, train ML models, run experiments, analyze results, and iterate through this process to make new scientific discoveries. At Meta, she also started and led the Tool Use team for Llama 3. This enabled Llama to use tools such as a search engine, Python interpreter, text-to-image models, and Wolfram Alpha, as well as zero-shot generalize to new tools at test time. Her research on LLM Agents enabled new products used by hundreds of millions

of users such as Meta AI, Data Analyst, AI Studio, and the Ads Business Agent. In 2021, she obtained her PhD in Computer Science from NYU, advised by Rob Fergus. Her focus was on deep reinforcement learning. During her PhD, she was fortunate to intern at DeepMind, Microsoft Research, and Facebook AI Research. Previously, she obtained a B.A. in Astrophysics from Princeton University, where she worked with Michael Strauss on theoretical cosmology and Eve Ostriker on supernovae simulations. In a past life (aka high school), she was fortunate to have the opportunity of competing in IPhO, IOAA, IAO, as well as national math and computer science olympiads.

Keynote Talk

Linguistic Agents that Learn from Interactions

Christopher Manning
Stanford University

Thursday, July 31, 2025 – Time: **12:00 – 12:30** – Room: **Room 1.61-62**

Abstract:

Bio: Christopher Manning is the inaugural Thomas M. Siebel Professor in Machine Learning in the Departments of Linguistics and Computer Science at Stanford University, a Founder and Associate Director of the Stanford Institute for Human-Centered Artificial Intelligence (HAI), and was Director of the Stanford Artificial Intelligence Laboratory (SAIL) from 2018-2025. From 2010, Manning pioneered Natural Language Understanding and Inference using Deep Learning, with impactful research on sentiment analysis, paraphrase detection, the GloVe model of word vectors, attention, neural machine translation, question answering, self-supervised model pre-training, tree-recursive neural networks, machine reasoning, summarization, and dependency parsing, work for which he has received two ACL Test of Time Awards and the IEEE John von Neumann Medal (2024). He earlier led the development of empirical, probabilistic approaches to NLP, computational linguistics, and language understanding, defining and building theories and systems for natural language inference, syntactic parsing, machine translation, and multilingual language processing, work for which he won ACL, Coling, EMNLP, and CHI Best Paper Awards. In NLP education, Manning coauthored foundational textbooks on statistical NLP (Manning and Schütze 1999) and information retrieval (Manning, Raghavan, and Schütze, 2008), and his online CS224N Natural Language Processing with Deep Learning course videos have been watched by hundreds of thousands. In linguistics, Manning is a principal developer of Stanford Dependencies and Universal Dependencies, and has authored monographs on ergativity and complex predicates. He is the founder of the Stanford NLP group (@stanfordnlp) and was an early proponent of open source software in NLP with Stanford CoreNLP and Stanza. He is an ACM Fellow, a AAAI Fellow, and an ACL Fellow, and was President of the ACL in 2015. Manning earned a B.A. (Hons) from The Australian National University, a Ph.D. from Stanford in 1994, and an Honorary Doctorate from U. Amsterdam in 2023. He held faculty positions at Carnegie Mellon University and the University of Sydney before returning to Stanford.

Keynote Talk

Open Foundations for Computer-Use Agents

Tao Yu
University of Hong Kong

Thursday, July 31, 2025 – Time: **13:30 – 14:00** – Room: **Room 1.61-62**

Abstract: Recent advances in vision-language models (VLMs) have enabled AI agents to operate computers just as humans do. In this talk, I will present our open source efforts to scaling these agents through three key dimensions: data, methods, and evaluation. First, I will introduce how we leverage internet-scale instructional videos and human demonstrations via our AgentNet platform to build large-scale computer interaction datasets. I will then discuss our methods (SFT & RL) for training foundation models that ground natural language into interface actions. Finally, I will present Agent Arena, our open platform for scalable real-world evaluation through crowdsourced user computer interactions, and outline key directions for improving agent robustness and safety for real-world deployment.

Bio: Tao Yu is an Assistant Professor of Computer Science at The University of Hong Kong and a director of the XLANG Lab (as part of the HKU NLP Group). He spent one year in the UW NLP Group working with Noah Smith, Luke Zettlemoyer, and Mari Ostendorf. He completed his Ph.D. in Computer Science from Yale University, advised by Dragomir Radev and master's at Columbia University advised by Owen Rambow and Kathleen McKeown.
Tao has received the Google and Amazon faculty research awards (Google Research Scholar Award 2023, Amazon Research Award 2022). His main research interest is in Natural Language Processing. His research aims to develop embodied AI agents that empower users to use language to interact with digital and physical environments to carry out real-world tasks. Such systems need to ground language and perception into code and actions executable in the corresponding embodied environment, helping people perform data science, control computers, and collaborate with robots.

Keynote Talk

Augmenting Human Cognition with AI Agents that Use Computers

Yu Su
Ohio State University

Thursday, July 31, 2025 – Time: **14:00 – 14:30** – Room: **Room 1.61-62**

Abstract: Human cognition is the most advanced form of intelligence to date, but it is not without limitations. With a fixed capacity, there are limits on how fast we can perceive, think, and act. A new generation of AI agents, powered by multimodal large language models (LLMs), is emerging that can operate in the digital world as humans do. By automating tedious legwork, such agents provide a compelling way to augment human cognition; they enable us to focus our limited cognitive capacity on things that matter more, such as critical decisions, oversight, and creativity. In this talk, I will review work from my group on computer-use agents, ranging from early text-based agents to multimodal agents with a human-like embodiment and long-horizon agentic search systems that take hundreds of steps for complex tasks. I will conclude the talk with exciting future directions.

Bio: Yu Su is an Associate Professor at the Department of Computer Science and Engineering, The Ohio State University, where he co-direct the OSU NLP group, co-lead the Foundational AI team in the ICICLE AI Institute and lead the Machine Learning Foundations team in the Imageomics Institute. He obtained his PhD from University of California, Santa Barbara and his bachelor's degree from Tsinghua University, both in Computer Science. He also spent some fun time as a researcher at Microsoft Semantic Machines. He is a 2025 Sloan Research Fellow and received several best/outstanding paper awards from CVPR and ACL.
He is broadly interested in artificial intelligence, with a primary interest in the role of language as a vehicle for reasoning and communication. These days, he spends much of my time thinking about language agents [blog, tutorial], an emerging class of AI agents characterized by their language understanding and production capabilities.

Keynote Talk

Inducing Functions to Improve LLM Agents

Daniel Fried
Carnegie Mellon University

Thursday, July 31, 2025 – Time: **14:30 – 15:00** – Room: **Room 1.61-62**

Abstract: Programs provide a structured, reusable, and verifiable means for people to carry out digital tasks. We show that LLM-based agents also benefit from generating code, executing that code in an environment, and abstracting functions from correct code. We present tool induction methods that build libraries of reusable functions online as the agent interacts with the environment. Our methods allow agents to carry out tasks more accurately and efficiently in grounded environments including performing tasks on the web and answering questions about structured data and images. We also find that induced tools make agent trajectories easier to verify for people and generalize well to complex tasks with shared sub-structure.

Bio: Daniel Fried is an assistant professor at the Language Technologies Institute in the School of Computer Science at Carnegie Mellon University, working on natural language processing. He is also a research scientist at Meta.
His work focuses on enabling people to use language to interact with computers to carry out useful tasks in the world. One recurring theme in his work is pragmatics: viewing language as an action that people take in context to affect their communicative partners. He is excited about domains where computers can complement human abilities. Recently, he has been focusing on code generation, aiming to make programming more communicative.
Previously, he was a postdoc at FAIR Seattle and the University of Washington. He completed a PhD at UC Berkeley in the NLP Group and the Berkeley AI Research Lab, an M.Phil. at the Cambridge Computer Laboratory and a B.S. at the University of Arizona.

Table of Contents

Prompt-based Personality Profiling: Reinforcement Learning for Relevance Filtering
Jan Hofmann, Cornelia Sindermann and Roman Klinger 1

DFLOW: Diverse Dialogue Flow Simulation with Large Language Models
Wanyu Du, Song Feng, James Gung, Lijia Sun, Yi Zhang, Saab Mansour and Yanjun Qi 17

CAMPHOR: Collaborative Agents for Multi-input Planning and High-Order Reasoning On Device
Yicheng Fu, Raviteja Anantha and Jianpeng Cheng 33

A Multi-AI Agent System for Autonomous Optimization of Agentic AI Solutions via Iterative Refinement and LLM-Driven Feedback Loops
Kamer Ali Yuksel, Thiago Castro Ferreira, Mohamed Al-Badrashiny and Hassan Sawaf 52

The Art of Tool Interface Design
Yunnan Wu, Qile P. Chen, Deshank Baranwal, Jinlong Zhou and Jian Yuan 63

AID-Agent: An LLM-Agent for Advanced Extraction and Integration of Documents
Bin Li, Jannis Conen and Felix Aller 80

Hidden Forms: A Dataset to Fill Masked Interfaces from Language Commands
Anirudh Sundar, Christopher Gordon Richardson, William Gay, Benjamin Reichman and Larry Heck 89

Do Large Language Models Learn Human-Like Strategic Preferences?
Jesse Roberts, Kyle Moore and Douglas Fisher 97

Inherent and emergent liability issues in LLM-based agentic systems: a principal-agent perspective
Garry A. Gabison and R. Patrick Xian 109

Positive Experience Reflection for Agents in Interactive Text Environments
Philip Lippmann, Matthijs T. J. Spaan and Jie Yang 131

PAARS: Persona Aligned Agentic Retail Shoppers
Saab Mansour, Leonardo Perelli, Lorenzo Mainetti, George Davidson and Stefano D'Amato . 143

Leveraging LLM-based sentiment analysis for portfolio optimization with proximal policy optimization
Kemal Kirtac and Guido Germano 160

Safe in Isolation, Dangerous Together: Agent-Driven Multi-Turn Decomposition Jailbreaks on LLMs
Devansh Srivastav and Xiao Zhang 170

ToolReflection: Improving Large Language Models for Real-World API Calls with Self-Generated Data
Gregory Polyakov, Ilseyar Alimova, Dmitry Abulkhanov, Ivan Sedykh, Andrey Bout, Sergey Nikolenko and Irina Piontkovskaya 184

Conditional Multi-Stage Failure Recovery for Embodied Agents
Youmna Farag, Svetlana Stoyanchev, Mohan Li, Simon Keizer and Rama Doddipatla 200

Snap Out of It: A Dual-Process Approach to Mitigating Overthinking in Language Model Reasoning
Ashish Pandian, Nelson Lojo, Wei Xun Lai and Jackson Lukas 228

A Conversational Agent Framework for Multimodal Knowledge Retrieval: A Case Study in FHWA Info-Highway Web Portal Queries
Sai Surya Gadiraju, Duoduo Liao and Zijie He 250

A Study on Leveraging Search and Self-Feedback for Agent Reasoning
Karthikeyan K, Michelle Yuan, Elman Mansimov, Katerina Margatina, Anurag Pratik, Daniele Bonadiman, Monica Sunkara, Yi Zhang and Yassine Benajiba 259

GitGoodBench: A Novel Benchmark For Evaluating Agentic Performance On Git
Tobias Lindenbauer, Egor Bogomolov and Yaroslav Zharov 272

TCQA^2: A Tiered Conversational Q&A Agent in Gaming
Ze Chen, Chengcheng Wei, Jiewen Zheng and Jiarong He 289

Oversight Structures for Agentic AI in Public-Sector Organizations
Chris Schmitz, Jonathan Rystrøm and Jan Batzner 298

Are You Sure You're Positive? Consolidating Chain-of-Thought Agents with Uncertainty Quantification for Aspect-Category Sentiment Analysis
Filippos Ventirozos, Peter A. Appleby and Matthew Shardlow 309

Bridging the Digital Divide: Empowering Elderly Smartphone Users with Intelligent and Human-Centered Design in Agemate
Liangliang Chen and Yongzhen Mu 327

Decentralized Low-Rank Fine-Tuning of Large Language Models
Sajjad Ghiasvand, Mahnoosh Alizadeh and Ramtin Pedarsani 334

Measuring temporal effects of agent knowledge by date-controlled tool use
R. Patrick Xian, Qiming Cui, Stefan Bauer and Reza Abbasi-Asl 346

VisTRA: Visual Tool-use Reasoning Analyzer for Small Object Visual Question Answering
Hiroaki Sugiyama, Ko Koga and Toshifumi Nishijima 356

StateAct: Enhancing LLM Base Agents via Self-prompting and State-tracking
Nikolai Rozanov and Marek Rei 367

DIAMOND: An LLM-Driven Agent for Context-Aware Baseball Highlight Summarization
Jeonghun Kang, Soonmok Kwon, Joonseok Lee and Byung-Hak Kim 386

RL + Transformer = A General-Purpose Problem Solver
Micah Rentschler and Jesse Roberts 401

From Knowledge to Noise: CTIM-Rover and the Pitfalls of Episodic Memory in Software Engineering Agents
Tobias Lindenbauer, Georg Groh and Hinrich Schuetze 411

FrontierScience Bench: Evaluating AI Research Capabilities in LLMs
Matthew Li, Santiago Torres-Garcia, Shayan Halder, Phani Kuppa, Sean O'Brien, Vasu Sharma, Kevin Zhu and Sunishchal Dev 428

The Power of Simplicity in LLM-Based Event Forecasting
Meiru Zhang, Auss Abbood, Zaiqiao Meng and Nigel Collier 454

Weight-of-Thought Reasoning: Exploring Neural Network Weights for Enhanced LLM Reasoning
Saif Punjwani and Larry Heck 471

Preventing Rogue Agents Improves Multi-Agent Collaboration
Ohav Barbi, Ori Yoran and Mor Geva 486