Generative AI and Retrieval-Augmented Generation (RAG) Systems for Enterprise

Anbang Xu
NVIDIA
Santa Clara, California, United States
xabang@gmail.com

Tan Yu
NVIDIA
Santa Clara, California, United States
tayu@nvidia.com

Min Du
NVIDIA
Santa Clara, California, United States
min.du.email@gmail.com

Pritam Gundecha
NVIDIA
Santa Clara, California, United States
pritam.gundecha@gmail.com

Yufan Guo
Amazon
Seattle, Washington, United States
yufan.guo@gmail.com

Xinliang Zhu
Amazon
Palo Alto, California, United States
xlzhu@amazon.com

May Wang
Santa Clara, California, United States
maywang@paloaltonetworks.com

Ping Li
VECML
Seattle, Washington, United States
xlzhu@amazon.com

Xinyun Chen
Google Brain
Mountain View, California, United States
xinyunchen@google.com

ABSTRACT

This workshop introduces generative AI applications for enterprise, with a focus on retrieval-augmented generation (RAG) systems. Generative AI is a field of artificial intelligence that can create new content and solve complex problems. RAG systems are a novel generative AI technique that combines information retrieval with text generation to generate rich and diverse responses. RAG systems can leverage enterprise data, which is often specific, structured, and dynamic, to provide customized solutions for various domains. However, enterprise data also poses challenges such as scalability, security, and data quality. This workshop convenes researchers and practitioners to explore RAG and other generative AI systems in real-world enterprise scenarios, fostering knowledge exchange, collaboration, and identification of future directions. Relevant to the CIKM community, the workshop intersects with core areas of data science and machine learning, offering potential benefits across various domains.

CCS CONCEPTS

• **Applied computing** → **Document searching**; • **Computing methodologies** → *Machine learning algorithms.*

ACM Reference Format:
Anbang Xu, Tan Yu, Min Du, Pritam Gundecha, Yufan Guo, Xinliang Zhu, May Wang, Ping Li, and Xinyun Chen. 2024. Generative AI and Retrieval-Augmented Generation (RAG) Systems for Enterprise. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24), October 21–25, 2024, Boise, ID, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3627673.3680117

1 MOTIVATION

Generative AI (GenAI) represented by large language models (LLMs) is revolutionizing the way we approach problem-solving and content creation in various domains. With its ability to interact with human in natural language, as well as its enormous knowledge base trained from gigantic volume of data, it holds immense potential for enterprise applications. In fact, almost all major companies now have dedicated teams focusing on developing GenAI solutions to better serve their customers and for internal applications. However, the unique characteristics of enterprise data, such as its specificity, structure, and dynamic nature, pose distinct challenges. Retrieval-Augmented Generation (RAG) systems, a subset of generative AI, offer a promising solution by combining information retrieval with text generation to create rich, contextually relevant responses. These systems can harness enterprise data to provide tailored solutions, addressing the challenges of scalability, security, and data quality. Moreover, enterprise applications pose additional challenges such as scalability, security, reliability, and compliance to GenAI solutions.

This workshop aims to bring together academic researchers and industrial practitioners who are interested in building GenAI solutions for enterprise AI, with a special focus on RAG systems. The workshop will provide a platform for sharing the latest research advances, practical experiences, and real-world challenges in this emerging field. The workshop will also foster collaboration and cross-disciplinary discussion among the participants, and identify future directions and opportunities. The workshop is highly relevant to the CIKM community, as it covers topics such as information retrieval, text and code generation, vector search, embeddings, representation learning, and reinforcement learning, which are all core areas of data science and machine learning. The workshop also aligns with the theme of “Data Science for Good”, as it seeks solutions that can improve the quality and efficiency of enterprise AI systems, and benefit various domains such as IT, HR, healthcare, education, finance, and e-commerce.

CIKM '24, October 21–25, 2024, Boise, ID, USA

ACM ISBN 979-8-4007-0436-9/24/10
https://doi.org/10.1145/3627673.3680117

Activity	Time Duration
Introduction	10 minutes
Invited Talk - I	20 mins
Invited Talk - II	20 mins
Contributed Talks - 4 papers (15 minutes each)	60 mins
Break and Poster Session	30 minutes
Invited Talk - III	20 mins
Panel Discussion	40 mins

Table 1: Workshop schedule.

Organizing this workshop in conjunction with CIKM 2024 is timely and crucial. The rapid advancements in generative AI and the growing interest in RAG systems necessitate a dedicated platform for discussion and exploration. As enterprises are increasingly harnessing GenAI solutions, the need for innovative solutions and best-practice sharing is more pressing than ever. This workshop will provide an opportunity to address these needs, showcase the latest developments, and set the agenda for future research in generative AI and RAG systems for enterprise applications.

2 WORKSHOP DETAILS

2.1 Workshop Agenda

We expect this workshop will be a half-day event. The activities include: three invited talks, four oral presentations from the accepted submissions, a poster session and a panel discussion. Table1 illustrates the detailed schedule.

2.2 Organizers

- **Anbang Xu** is Senior Manager at NVIDIA. He leads a Machine Learning team to develop enterprise AI solutions. His research is a mix of Applied Machine Learning and HCI. He is the Associate Editor of ACM Transactions on Interactive Intelligent Systems. He has published 50+ research articles and 25+ patents and received 3,300+ citations. He received his Ph.D. in Computer Science from the University of Illinois at Urbana-Champaign.
- **Tan Yu** is a Senior Staff Machine Learning Engineer in the Enterprise AI team at NVIDIA. His research focuses on information retrieval, recommendation system and multi-modal understanding. He has published 40+ research articles in top-tier conferences and journals including SIGKDD, CIKM, SIGIR, CVPR, ICCV, IJCV, ICLR, ECCV, EMNLP, NAACL, AAAI, IJCAI. He was a Staff Machine Learning Engineer in TikTok Advertisements Recommendation and a Senior Research Scientist in Baidu USA.
- **Min Du** is a Senior Staff Machine Learning Engineer in the Enterprise AI team at NVIDIA. She received her Ph.D. degree from University of Utah, followed by a Postdoc as UC Berkeley. Her research in AI, security and privacy has been published in various top venues with 3,000 citations. She has served on the Program Committee for multiple workshops in various top conferences including ICLR and IEEE S&P. She received a best poster award at SoCC 2017, and a best paper award at ICPE 2022. She was named as a Rising Star in EECS in 2019.
- **Pritam Gundecha** is a Senior Staff Data Scientist at NVIDIA for last two years. He obtained his PhD in Computer Science at the Arizona State University in 2015. Before he joined NVIDIA, he worked at HP Enterprises and IBM Research as a research staff member. He has a strong background in researching, developing and deploying large-scale deep-learning solutions in production. His research interests are in Generative AI and its applications, Deep Learning, Machine learning and natural language processing. His well cited publications include book, encyclopedia entries as well as patents, conference and journal papers.
- **Yufan Guo** is an Applied Science and Engineering Manager at Amazon, building foundation models to improve online shopping experience. Prior to Amazon, she was a Research Staff Member at IBM, where her work combined natural language processing, medical image analysis, healthcare informatics, and insights from cognitive computing. Yufan holds a PhD in Computation, Cognition and Language and an MPhil in Computer Speech, Text and Internet Technology from the University of Cambridge.
- **Xinliang Zhu** is a senior applied scientist from the Visual Search and Augmented Reality team at Amazon.com. Since joining Amazon post-graduation, Xinliang has shown a track record in keeping improving the Amazon visual search results quality through developing solutions that synergize data and algorithms. Recently, he led the effort of developing and launching multimodal search product to Amazon worldwide customers. Inside Amazon, Xinliang actively contributed as session chairs for an internal annual computer vision conference. He got his PhD degree in Computer Science from UT Arlington in 2020. His work had been accepted by CVPR/ECCV/MICCAI/AAAI and accumulated 1400 citations.
- **May Wang** is the CTO of IoT Security at Palo Alto Networks, a multinational cybersecurity company. She is leading AI initiatives across multiple security domains. She co-founded Zingbox which was acquired by Palo Alto Networks in 2019 for its AI-based IoT security solutions. Dr. Wang holds a Ph.D. in electrical engineering from Stanford University. She received numerous awards, including the prestigious 2023 Women in AI Entrepreneur award from VentureBeat.
- **Ping Li** is CEO of VecML. He is a PhD from Stanford, a former Distinguished Engineer at LinkedIn, and a former Chief Architect at Baidu (with teams in Bellevue, Sunnyvale, Beijing). Ping Li was also a professor at Cornell and Rutgers Universities and was listed by www.csrankings.org as a top researcher in machine learning and information retrieval.
- **Xinyun Chen** is a Senior Research Scientist at Google DeepMind. She received her Ph.D. from UC Berkeley. Her research focuses on large language models, reasoning, code generation, and trustworthy machine learning. Her work SpreadsheetCoder for spreadsheet formula prediction was integrated into Google Sheets, and her work AlphaCode was featured as the front cover in Science Magazine. She has co-organized 14 workshops and tutorials at ICLR, ICML, CVPR,

Speaker	Topic
Rama Akkiraju, VP AI/ML for IT, NVIDIA	Enterprise RAG Systems
May Wang, CTO, Internet of Things Security, Palo Alto Networks	Generative AI Systems for Enterprise Use Cases
Sravan Bodapati, Principal Scientist and Sr. Manager, AGI Foundation Models, Amazon	Generative AI Systems for Enterprise Use Cases

Table 2: Invited speakers

ECCV, ICCV, AAAI, etc. In particular, she co-organized the LLM agents workshop at ICLR 2024, and she is a co-chair of AISec 2022 and 2023.

2.3 Invited speakers (tentative)

We list the invited speakers and the topics of their talks in Table 2. To be specific, our invited speakers include:

- **Rama Akkiraju**, VP AI/ML for IT, NVIDIA
- **May Wang**, CTO, Internet of Things Security, Palo Alto Networks
- **Jacob Liberman**, Sr Manager, Enterprise GenAI Product Management, NVIDIA
- **Guang Cheng**, Amazon Scholar/Professor of Statistics and Data Science at UCLA

2.4 Panellists (tentative)

- **Jasmine Jaksic**, Senior Director, NVIDIA
- **Aaron Isaksen**, VP of AI/ML in Cortex, Palo Alto Networks
- **Bo Li**, Associate Professor, University of Chicago
- **Jaewon Yang**, Principal Machine Learning Engineer, Nextdoor
- **Sravan Bodapati**, Principal Applied Scientist, Amazon

2.5 Program committee (tentative)

Our tentative program committee members are listed as follows.

- Anbang Xu, PhD, NVIDIA, USA
- Tan Yu, PhD, NVIDIA, USA
- Min Du, PhD, NVIDIA, USA
- Pritam Gundecha, PhD, NVIDIA, USA
- Xinliang Zhu, PhD, Amazon, USA
- Xinyun Chen, PhD, Google DeepMind, USA
- Yanqing Peng, PhD, Meta, USA
- Yan Zheng, PhD, Visa Research, USA
- Guanhua Wang, PhD, Microsoft Research, USA

3 TENTATIVE CFP

3.1 Topics of interest

We seek contributions on all aspects related to GenAI usages in Enterprise AI.

The topics of interest include, but are not limited to:

- Enterprise AI dataset and collection techniques
- RAG systems for specific enterprise domains or tasks
- Retrieval strategies and models for RAG systems
- Text-to-SQL design and solutions
- Foundation model pre-training, prompting and fine-tuning
- Hallucination prevention
- Multi-modal and multi-lingual
- Evaluation metrics
- Scalability and efficiency
- Security and privacy aspects
- Compliance and ethical issues
- User feedback and interaction mechanisms
- Case studies and best practices

3.2 Submission instructions

All submissions should follow the CIKM 2024 format and guidelines.

We accept two types of paper submissions:

Track 1: Full length papers. Submissions in this track can be up to 6 pages, plus any pages of additional references and appendices. Authors should be noted that reviewers are not required to read appendices.

Track 2: Extended abstract. Submission for this track can be up to 2 pages, plus any pages of additional references and appendices. Authors should be noted that reviewers are not required to read appendices. For this track, we encourage submissions that describe visionary ideas, preliminary results, controversial findings, or experience sharing. Acceptable material includes work that has already been submitted or published. Authors are responsible for ensuring compliance with the policies of other venues.

All papers should be submitted in a single PDF format through Microsoft CMT website. For dataset and code submissions, authors should include an external link in the submitted PDF.

3.3 Presentation format

Accepted papers will be presented at the workshop in the form of either talk or poster, depending on their received review score and novelty. Remote presentation is allowed.

4 VENUES TO ADVERTISE

First of all, all our organizers are GenAI leaders and influencers from different companies. They are committed to promote this event through their internal company channels, as well as their personal networks.

Additionally, below is a list of possible venues (e.g., mailing lists, social media) to advertise the proposed workshop:

- Linkedin
- Twitter
- ml-news@googlegroups.com
- NVIDIA News

5 EXPECTED ATTENDEES/SUBMISSIONS

This workshop is designed for researchers, professionals, and students specializing in information retrieval (IR), generative artificial intelligence (GenAI), large language model (LLM), and the enterprise applications. Attendees will have the opportunity to deepen their understanding of empowering enterprise products by the state-of-the-art LLM models.

We expect around 50 submissions and 60 attendees. The justification is as below:

- Our organizers are from 5 different companies, where GenAI is actively researched and developed for different product domains. We expect an average of 5 submissions from each company. Given the focus and popularity of GenAI in enterprise, we expect an additional of 15 submissions from other companies as well as 10 submissions from academia.
- Most companies have a budget for their employees to attend at least one conference each year. Our workshop is one of the most relevant events for industry practitioners in this area. We allow for remote attendance, which gives more flexibility. Academic researchers who are looking for real-world applications or industry jobs will also be attracted. We expect 35 attendees from industry and 25 attendees from academia.