Workshop on Large Language Models' Interpretability and Trustworthiness (LLMIT)

Tulika Saha
Department of Computer Science
University of Liverpool, United Kingdom
sahatulika15@gmail.com

Debasis Ganguly
School of Computing Science
University of Glasgow, United Kingdom
debasis.ganguly@glasgow.ac.uk

Sriparna Saha
Department of Computer Science
Indian Institute of Technology Patna, India
sriparna@iitp.ac.in

Prasenjit Mitra
L3S Research Center, Germany
& The Pennsylvania State University, U.S.A.
mitra@l3s.de

ABSTRACT

Large language models (LLMs), when scaled from millions to billions of parameters, have been demonstrated to exhibit the so-called 'emergence' effect, in that they are not only able to produce semantically correct and coherent text, but are also able to adapt themselves surprisingly well with small changes in contexts supplied as inputs (commonly called prompts). Despite producing semantically coherent and potentially relevant text for a given context, LLMs are vulnerable to yield incorrect information. This misinformation generation, or the so-called hallucination problem of an LLM, gets worse when an adversary manipulates the prompts to their own advantage, e.g., generating false propaganda to disrupt communal harmony, generating false information to trap consumers with target consumables etc. Not only does the consumption of an LLM-generated hallucinated content by humans pose societal threats, such misinformation, when used as prompts, may lead to detrimental effects for in-context learning (also known as few-shot prompt learning). With reference to the above-mentioned problems of LLM usage, we argue that it is necessary to foster research on topics related to not only identifying misinformation from LLM-generated content, but also to mitigate the propagation effects of this generated misinformation on downstream predictive tasks thus leading to more robust and effective leveraging in-context learning.

CCS CONCEPTS

• **Information systems** → **Information retrieval**; • **Computing methodologies** → **Machine learning**; **Natural language processing**.

KEYWORDS

Large Language Model, Trustworthiness, Interpretability, In-context Learning, Explainability

CIKM '23, October 21–25, 2023, Birmingham, United Kingdom

ACM ISBN 979-8-4007-0124-5/23/10…$15.00
https://doi.org/10.1145/3583780.3615311

ACM Reference Format:
Tulika Saha, Debasis Ganguly, Sriparna Saha, and Prasenjit Mitra. 2023. Workshop on Large Language Models' Interpretability and Trustworthiness (LLMIT). In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23), October 21–25, 2023, Birmingham, United Kingdom.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3583780.3615311

1 MOTIVATION OF THE WORKSHOP

Research on Large Language Models (LLMs) is expanding in scope and yielding significant scientific advancements at a rapid rate. LLMs, when scaled from millions to billions of parameters, have been demonstrated to exhibit the so-called 'emergence' effect [15], in that they are not only able to produce semantically correct and coherent text, but are also able to adapt themselves surprisingly well with small changes in contexts supplied as inputs (commonly called prompts) [1].

Motivated by this generic task-independent *knowledge representation* ability of the LLMs [25], several threads of research has explored their usage in a wide range of tasks such as text classification [1, 20], conversational systems [8, 18, 19], information retrieval [11, 14, 16], recommendation systems [13], and even to assist manual assessments [23].

Despite the benefits of LLMs, there are two major concerns that the community should be aware of, and foster research towards addressing them. These concerns are listed below.

(1) **LLMs hallucinate** [4, 9]: LLMs can often generate text, which although being coherent and apparently correct, actually constitutes of information that is not true, e.g., events that never happened, or entities that are not real.
(2) **Poor adversarial robustness**: LLMs are likely exhibit a poor adversarial robustness, i.e., it is relatively easy for an adversary to manipulate the prompts to their own advantage, e.g., generating false propaganda to disrupt communal harmony, generating false information to trap consumers with target consumables etc. Not only does the consumption of an LLM-generated hallucinated content by humans pose societal threats, such misinformation, when used as prompts, may lead to changing the behaviour of a predictive system to an adversary's advantage.

To alleviate the above-mentioned problems, it is critical that researchers work towards a responsible and trustworthy innovation [3]. For instance, identifying the source documents which the decoding phase of an LLM largely utilised to generate an output text can potentially help a user better understand the content, thus possibly contributing towards developing a trust towards the model [2, 11].

While explainable AI has gained popularity in terms of providing feature attributions for supervised models both locally and globally, i.e., at the level of individual instances and the overall model [12, 17, 24], such investigation also needs to be conducted for the 0-shot and few-shot generation phase of LLMs. Particularly, for in-context (few-shot) learning of LLMs, i) attribution of documents during decoding (i.e., which documents were likely used to generate the decoded text) [2], ii) attribution over discrete terms of an encoder state [11], iii) attribution over the set of prompts used for k-shot task-specific tuning, i.e., what parts of a prompt are mainly responsible for the prediction (in terms of the ⟨mask⟩ token class probabilities during decoding) etc. can provide a user and a practitioner useful insights about a model's behaviour.

The need for such explainable and trustworthy in-context learning (prompt learning) means that it is timely to organize a workshop to bring together experts working in the area of responsible AI to further advance the state of the art in trustworthy use of LLMs for various downstream tasks. The broad range of researchers with relevant expertise associated with CIKM 2023 makes this an ideal venue for this workshop.

Relevant active areas in LLMs include misinformation detection, in-context learning, detection and mitigation of biases using prompts, alleviating adversarial attacks etc. for several downstream tasks to aid the goal of responsible AI [22]. On the other hand, these responsible AI objectives are often viewed independently to each other, which is a significant constraint because it is necessary to comprehend the interplay and/or conflict between them. The purpose of this workshop is to bring together researchers working on those distinct yet connected areas, as well as their overlap, in order to move towards a more thorough understanding of trustworthy use of LLMs.

2 WORKSHOP DETAILS

We name our workshop '**Large Language Models' Interpretability and Trustworthiness**', abbreviated as LLMIT (pronounced as 'limit'). Because of the widespread interest on LLMs, we believe that we will witness an enthusiastic participation. Thus, we propose LLMIT to be a full-day workshop in a hybrid format enabling engagement of both in-person and online participants. The proposed activities of the workshop are as follows:

(1) **Invited speakers:** We propose to include one or two invited speakers who have an established track record in advocating for or working on responsible AI focused on LLMs.
(2) **Paper presentations:** Depending on the number of accepted submissions, papers will be presented in oral and/or poster sessions.
(3) **Panel Session:** We plan to organize a panel session comprised of the invited speakers to explore perspectives on interpretability of LLMs, and explore possible focus areas for future research.
Breakout session: We plan to organize a breakout session, where we plan to pair up researchers in small groups for a more direct interchange of ideas related to specific subtopics of interests on interpretation of LLMs.

3 TOPICS FOR THE CALL FOR PAPERS

The broad-level topics that are of interest to the LLMIT workshop include work directed towards LLM interpretation and their applications towards mitigation of their hallucination effects. However, to encourage a wider participation, we solicit submissions on other closely related topics as well. More explicitly, the following are some of the research topics of interest.

(1) **Misinformation detection**: It is a widely known fact that the text generated by LLMs are often of hallucinatory in nature [4, 9]. Consequently, mapping back to the original documents (attribution at document level [2]) which were likely responsible in affecting the decoding path to generate the text for a downstream task can throw insights on an LLM's 0-shot or few-shot (in-context) task-specific abstraction of the input data.
(2) **Prompt or In-Context Explanations** of the LLM-generated content in terms of attention-based or counter-factual explanations of the prompts, i.e., which parts of the prompts are more important in determining the class probabilities (via generation of class-specific sets of words by the decoder).
(3) **Linking LLM-generated answers to knowledge bases** or use knowledge-bases to formulate template-driven prompts.
(4) **LLMs for generating weak labels** for various applications, or to generate simulated data (silver-standard ground-truth) to reduce annotation effort [23].
(5) **Adaptive In-context learning for LLMs**, i.e., work towards developing a transparent LLM-based in-context learning model that explains the different choices employed in prompt learning, which may include -
   - What similarity metric to employ for generating the few-shot examples, e.g., sparse vectors, dense embedding, task-specific dense embedding etc.
   - How many few shot examples to use.
   - Explore combination (apriori and post-hoc) of LLM-based in-context learning with supervised parametric models.
(6) **Fair Predictions with LLMs**, i.e., mitigate the detrimental effects of biased responses with suitable prompts.
(7) **Adversarial Robustness of LLMs**, i.e., optimise the robustness of LLM-based in context learning to adversarial attacks based on prompt injections.
(8) **LLM-driven in-context learning for search and recommendation**, i.e., explore the potential of LLMs for personalised search and recommendation. Since LLMs have been shown to work well with small quantities of task-specific training data, they can potentially be used to improve the effectiveness of personalized search and recommendation.
(9) **Multi-modal LLMs**, involving exploration on the topics related to the interpretability and trustworthiness of LLMs in the particular context of multimodal predictive tasks, e.g., visual question answering etc.
(10) **Ethical concerns of LLMs**, i.e., research on topics related to a responsible use of LLMs and their socio-economic implications.

4 TENTATIVE PROGRAM COMMITTEE MEMBERS

A tentative list of members, who are supportive of the workshop idea and have tentatively agreed to volunteer in the program committee, is listed below (upon acceptance of the workshop proposal, we will add more pertinent people to this list).

- Rishiraj Saharoy, Max Planck Institute for Informatics, Germany.
- Sean MacAvaney, University of Glasgow, United Kingdom.
- Claudia Hauff, Delft University of Technology, Netherlands.
- Sumit Bhatia, IBM Research, India.
- Francesca Bonnin, IBM Research, Dublin.
- Charles Clarke, University of Waterloo, Canada.
- Xi Wang, University College London.
- Nicola Tonellotto, University of Pisa.

5 RELATED WORKSHOPS

There exists a small number of other workshops which are based on responsible AI as its core theme - the most recent one being "**TrustNLP: Workshop on Trustworthy Natural Language Processing**"[1] which was held consecutively in 2021, 2022 and is currently colocated with ACL 2023 for its third edition. This workshop differs from LLMIT in terms of its scope in that it encompasses a more generic theme of explainability and trustworthiness of any predictive model. However, our proposed workshop explicitly focuses on investigating these questions for the specific context of LLMs.

Another related workshop is "**ExplainAble Recommendation and Search**"[2], the main theme of which was to explore the issues related to the explainability and trustworthiness of search and recommender systems with a particular emphasis on the user behaviour. Our workshop, explores these topics related to NLP, IR and multimodal predictive models but restricted to the speific context of LLMs.

6 ORGANIZERS

(1) **Dr. Tulika Saha** (corresponding author) is a Lecturer of Computer Science in the University of Liverpool, United Kingdom (UK). Her current research interests include ML, DL, NLP typically Dialogue Systems, AI for Social Good, Social Media Analysis etc. She was a postdoctoral research fellow at the National Centre for Text Mining, University of Manchester, UK. Previously she earned her Ph.D. from Indian Institute of Technology Patna, India. Her research articles are published in top-tier conferences such as ACL, ACM SIGIR etc. and several peer-reviewed journals. She has organized several workshops/symposium namely NLP for Social Good 2023, AICAI 2023 and presented tutorials in ECIR [21], InterSpeech, 2023 etc.

(2) **Dr. Debasis Ganguly** is a Lecturer (Assistant Professor), School of Computing Science, University of Glasgow, Glasgow, Scotland. Formerly, he was a research staff member at IBM Research Europe, Dublin, Ireland. Generally speaking, his research activities span topics on IR and NLP. More specifically, he is interested in semantic search, neural retrieval models, explainable search and recommendation, fair and trustworthy search, and privacy preserving AI. Apart from this, he is interested in automatically constructing knowledge bases from legal documents for structured and explainable search. He is a part of the organization committee of the Symposium on Artificial Intelligence and Law (SAIL). His workshop organizational experience includes PASIR at CIKM'22 [10], SUD at WSDM'21 [5], SMERP-2018 at WWW'18 [7], and SMERP-2017 at ECIR'17 [6].

(3) **Dr. Sriparna Saha** is currently serving as an Associate Professor (h5-index:37, total citations: 7127 as per Google Scholar), in Computer Science and Engineering, Indian Institute of Technology Patna, India. Her current research interests include ML, DL, NLP, AI for Social Good, and Information Retrieval. She has published more than 400 papers in reputed journals (IEEE and ACM Transactions) and conferences including ACL, SIGIR, AAAI, EMNLP, ECIR, COLING, ACM MM, etc. She was one of the special session organizers of ICONIP 2021 on the topic of "Smart Home Technologies & Services for the Wellbeing and Sustainability of Society" and in IEEE SSCI 2021 on the topic of "Computational Intelligence for Natural Language Processing". She has also delivered tutorials in WCCI 2020, ICONIP 2022, ECIR 2023, InterSpeech 2023.

(4) **Prasenjit Mitra** is a Professor at The Pennsylvania State University and a visiting Professor at the L3S Center at the Leibniz University at Hannover, Germany. He obtained his Ph.D. from Stanford University in 2003 in Electrical Engineering and has been at Penn State since. His research interests are in artificial intelligence, applied machine learning, natural language processing, etc. His research has been supported by the NSF CAREER award, the DoE, DoD, Microsoft Research, Raytheon, Lockheed Martin, Dow Chemicals, McDonnell Foundation, etc. His has published over 200 peer-reviewed papers at top conferences and journals, supervised or co-supervised 15-20 Ph.D. dissertations; his work has been widely cited (h-index 60) and over 12,500 citations. Along with his co-authors, he has won the test of time award at the IEEE VIS and a best paper award at ISCRAM, etc. He has been the co-chair of several workshops, including a workshop previously collocated with CIKM. They are listed below:
   - Program Chair, Big-O(Q)'15: Workshop on Big-Graphs Online Querying in VLDB'15: the 41st International Conference on Very Large Databases (2015).
   - Program Chair, WIDM'12: The 12th International Workshop on Web Information and Data Management in CIKM'12: the 21st ACM International Conference on Information and Knowledge Management. (2012).
   - Program Co-Chair, WIDM'09: The 11th International Workshop on Web Information and Data Management in CIKM'09: the 18th ACM International Conference on Information and Knowledge Management. (2009).
   - Program Co-Chair, SNAKDD'09: The 2nd International Workshop on Social Network Mining and Analysis in KDD'08: the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. (2008).

[1] https://trustnlpworkshop.github.io/
[2] https://ears2019.github.io/

- Lead Program Co-Chair, CIMS'07: The 1st Workshop on CyberInfrastructure: Information Management in eScience (CIMS) in CIKM'07: the16th ACM International Conference on Information and Knowledge Management. (2007).

REFERENCES

[1] Simran Arora, Avanika Narayan, Mayee F. Chen, Laurel Orr, Neel Guha, Kush Bhatia, Ines Chami, Frederic Sala, and Christopher Ré. 2022. Ask Me Anything: A simple strategy for prompting language models. arXiv:2210.02441 [cs.CL]

[2] Bernd Bohnet, Vinh Q. Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, Kai Hui, Tom Kwiatkowski, Ji Ma, Jianmo Ni, Lierni Sestorain Saralegui, Tal Schuster, William W. Cohen, Michael Collins, Dipanjan Das, Donald Metzler, Slav Petrov, and Kellie Webster. 2023. Attributed Question Answering: Evaluation and Modeling for Attributed Large Language Models. arXiv:2212.08037 [cs.CL]

[3] Pin-Yu Chen and Chaowei Xiao. 2023. Trustworthy AI in the Era of Foundation Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[4] Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. 2022. On the Origin of Hallucinations in Conversational Models: Is it the Datasets or the Models?. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Seattle, United States, 5271–5285. https://doi.org/10.18653/v1/2022.naacl-main.387

[5] Debasis Ganguly, Manisha Verma, Procheta Sen, Dipasree Pal, and Gareth J. F. Jones. 2021. Overview of the Supporting and Understanding of Conversational Dialogues (SUD) Workshop. In *WSDM*. ACM, 1163–1164.

[6] Saptarshi Ghosh, Kripabandhu Ghosh, Debasis Ganguly, Tanmoy Chakraborty, Gareth J. F. Jones, and Marie-Francine Moens. 2017. ECIR 2017 Workshop on Exploitation of Social Media for Emergency Relief and Preparedness (SMERP 2017). *SIGIR Forum* 51, 1 (2017), 36–41.

[7] Saptarshi Ghosh, Kripabandhu Ghosh, Debasis Ganguly, Tanmoy Chakraborty, Gareth J. F. Jones, and Marie-Francine Moens. 2018. Report on the Second Workshop on Exploitation of Social Media for Emergency Relief and Preparedness (SMERP 2018) at the Web Conference (WWW) 2018. *SIGIR Forum* 52, 2 (2018), 163–168.

[8] Raghav Jain, Tulika Saha, Souhitya Chakraborty, and Sriparna Saha. 2022. Domain infused conversational response generation for tutoring based virtual agent. In *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[9] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* 55, 12, Article 248 (mar 2023), 38 pages. https://doi.org/10.1145/3571730

[10] Gareth J. F. Jones, Procheta Sen, Debasis Ganguly, and Emine Yilmaz. 2022. Workshop on Proactive and Agent-Supported Information Retrieval (PASIR). In *CIKM*. ACM, 5167–5168.

[11] Minghan Li, Xueguang Ma, and Jimmy Lin. 2022. An Encoder Attribution Analysis for Dense Passage Retriever in Open-Domain Question Answering. In *Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022)*. Association for Computational Linguistics, Seattle, U.S.A., 1–11. https://doi.org/10.18653/v1/2022.trustnlp-1.1

[12] Ninareh Mehrabi, Umang Gupta, Fred Morstatter, Greg Ver Steeg, and Aram Galstyan. 2022. Attributing Fair Decisions with Attention Interventions. In *Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022)*. Association for Computational Linguistics, Seattle, U.S.A., 12–25. https://doi.org/10.18653/v1/2022.trustnlp-1.2

[13] Sheshera Mysore, Andrew McCallum, and Hamed Zamani. 2023. Large Language Model Augmented Narrative Driven Recommendations. arXiv:2306.02250 [cs.IR]

[14] Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernández Ábrego, Ji Ma, Vincent Y. Zhao, Yi Luan, Keith B. Hall, Ming-Wei Chang, and Yinfei Yang. 2021. Large Dual Encoders Are Generalizable Retrievers. arXiv:2112.07899 [cs.IR]

[15] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[16] Ronak Pradeep, Kai Hui, Jai Gupta, Adam D. Lelkes, Honglei Zhuang, Jimmy Lin, Donald Metzler, and Vinh Q. Tran. 2023. How Does Generative Retrieval Scale to Millions of Passages? arXiv:2305.11841 [cs.IR]

[17] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1135–1144.

[18] Tulika Saha, Vaibhav Gakhreja, Anindya Sundar Das, Souhitya Chakraborty, and Sriparna Saha. 2022. Towards motivational and empathetic response generation in online mental health support. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*. 2650–2656.

[19] Tulika Saha, Saichethan Reddy, Anindya Das, Sriparna Saha, and Pushpak Bhattacharyya. 2022. A Shoulder to Cry on: Towards A Motivational Virtual Assistant for Assuaging Mental Agony. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Seattle, United States, 2436–2449. https://doi.org/10.18653/v1/2022.naacl-main.174

[20] Tulika Saha, Saichethan Miriyala Reddy, Sriparna Saha, and Pushpak Bhattacharyya. 2022. Mental health disorder identification from motivational conversations. *IEEE Transactions on Computational Social Systems* (2022).

[21] Tulika Saha, Abhishek Tiwari, and Sriparna Saha. 2023. Trends and Overview: The Potential of Conversational Agents in Digital Health. In *Advances in Information Retrieval*, Jaap Kamps, Lorraine Goeuriot, Fabio Crestani, Maria Maistro, Hideo Joho, Brian Davis, Cathal Gurrin, Udo Kruschwitz, and Annalina Caputo (Eds.). Springer Nature Switzerland, Cham, 349–356.

[22] Boaz Shmueli, Jan Fell, Soumya Ray, and Lun-Wei Ku. 2021. Beyond Fair Pay: Ethical Implications of NLP Crowdsourcing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (Eds.). Association for Computational Linguistics, 3758–3769. https://doi.org/10.18653/v1/2021.naacl-main.295

[23] Veniamin Veselovsky, Manoel Horta Ribeiro, and Robert West. 2023. Artificial Artificial Artificial Intelligence: Crowd Workers Widely Use Large Language Models for Text Production Tasks. arXiv:2306.07899 [cs.CL]

[24] Hanjing Wang, Dhiraj Joshi, Shiqiang Wang, and Qiang Ji. 2023. Gradient-based Uncertainty Attribution for Explainable Bayesian Deep Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12044–12053.

[25] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. 2022. Taxonomy of risks posed by language models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 214–229.