OMMM 2025

Proceedings of the First Interdisciplinary Workshop on Observations of Misunderstood, Misguided and Malicious Use of Language Models

associated with

The 15th International Conference on Recent Advances in Natural Language Processing 2025

September 11th, 2025
Varna, Bulgaria

The First Interdisciplinary Workshop
Observations of Misunderstood, Misguided
and Malicious Use of Language Models
Associated with the International Conference
Recent Advances in Natural Language Processing 2025
PROCEEDINGS
Varna, Bulgaria
11th September 2025

ISBN 978-954-452-101-1

Designed by INCOMA Ltd.
Shoumen, BULGARIA

Message from the Program Chairs (TODO)

Welcome to the proceedings of the first edition of the Interdisciplinary Workshop on Observations of Misunderstood, Misguided and Malicious Use of Language Models: (OMMM 2025), hosted at the 15th Biennial Conference on Recent Advances in Natural Language Processing (RANLP 2025), in Varna, Bulgaria.

OMMM 2025 is a new endeavour with the purpose of drawing together communities studying the inappropriate and harmful uses of Large Language Models (LLMs). In particular, the organising comittee is made up of experts from natural language processing, human computer interaction and psychology. Through this venture we aim to share common perspectives on the capabilities, vulnerabilities and harmful applications of LLMs. Our aim is to foster a new community drawn from various disciplines within and beyond our own, which is focussed on the mitigation of potential harms from the ever increasing ubiquity of AI technology powered by LLMs.

The use of Large Language Models (LLMs) pervades scientific practices in multiple disciplines beyond the NLP/AI communities. Alongside benefits for productivity and discovery, widespread use often entails misuse due to misalignment of values, lack of knowledge, or, more rarely, malice. LLM misuse has the potential to cause real harm in a variety of settings. Through this workshop, we aim to gather researchers interested in identifying and mitigating inappropriate and harmful uses of LLMs. For the purposes of designing a programme and motivating submissions, we categorised the misuses of LLMs into three domains:

- **Misunderstood** usages: Misrepresentation, improper explanation, or opaqueness of LLMs. Including: The use of anthropomorphic language by or for LLMs; Attributions of consciousness to LLM agents; Interpretability of LLM outputs or decisions; and harms arising from overreliance or misplaced trust in LLMs.
- **Misguided** usages: Misapplication of LLMs where their utility is questionable or inappropriate. Including: underperformance and inappropriate applications; structural limitations and ethical considerations; and deployment without proper training or safeguards.
- **Malicious** usages: Use of LLMs for misinformation, plagiarism, and adversarial attacks. Including: Adversarial attacks, jailbreaking; Detection and watermarking of machine-generated content; Generation of misinformation or plagiarism; and bias mitigation and trust design.

This year, we received 13 submissions to the workshop. These submissions covered a variety of current topics of interest in line with the aims of the workshop. In particular, the organisers noted submissions on anthropmorphised descriptions of LLMs, including new datasets for identification of anthropmorphisation; Bias as applied to large language models and the downstream harmful effects; case studies including negative results where LLMs failed compared to traditional approaches; the detection of AI generated texts; and work on AI alignment.

All submissions were peer-reviewed by the members of the program committee which includes specialists drawn from NLP, Philosophy,, Psychology, AI Ethics, LLM Security, and Misinformation. The organisers provided a further meta-review for all submissions, summarising the outcomes and decision as well as offering additional feedback. Out of the 13 submissions to the workshop, 9 were accepted, 2 were rejected and a further 2 papers were accepted subject to improvements in line with reviewer feedback. Each PC member had no more than three assignments. The organisers were delighted to see so many papers submitted in line with the mission of the workshop, demonstrating the necessity of such an event and the nascent community surrounding it.

The workshop is held in-person, with online attendance for authors who were unable to attend. The program encompasses: An introductory session ran by the organisers covering the grand challenges of misunderstood, misguided and malicious use. The programme then consists of 3 sessions, one covering papers submitted that are relevant to each of the topics: 6 papers were presented in the first session on *Misunderstood Use*. 2 papers were presented in the second session on **Misguided Use**. Finally, 3 papers were presented in the third session on **Malicious Use**.

Each session was succeeded by a discussion session, culminating in a final discussion session to close the event. The organisers intend to use the results of the discussions to co-create with the participants a future publication on the grand challenges of LLM Misuse.

We would like to thank the members of the program committee for their timely help in reviewing the submissions and all the authors for submitting their papers to the workshop. We also thank the organisers of RANLP for hosting the workshop and their kind support in producing these proceedings. Additionally, our thanks go to those who maintain the ACL Anthology in which these proceedings appear.

OMMM Organizing Committee

Piotr Przybyła, Matthew Shardlow, Nanna Inie, Clara Colombatto

Organizing Committee

- Piotr Przybyła, Pompeu Fabra University and Institute of Computer Sciences, Polish Academy of Sciences
- Matthew Shardlow, Manchester Metropolitan University
- Nanna Inie, IT University of Copenhagen
- Clara Colombatto, University of Waterloo

Programme Committee

- Alina Wróblewska (Polish Academy of Sciences)
- Ashley Williams (Manchester Metropolitan University)
- Azadeh Mohammadi (University of Salford)
- Clara Colombatto (University of Waterloo)
- Dariusz Kalociński (Polish Academy of Sciences)
- Julia Struß (Fachhochschule Potsdam)
- Lev Tankelevitch (Microsoft Research)
- Leon Derczynski (NVIDIA)
- Marcos Zampieri (George Mason University)
- Matthew Shardlow (Manchester Metropolitan University)
- Nael B. Abu-Ghazaleh (University of California, Riverside)
- Nanna Inie (IT University of Copenhagen)
- Nhung T. H. Nguyen (Johnson & Johnson)
- Nishat Raihan (George Mason University)
- Oluwaseun Ajao (Manchester Metropolitan University)
- Peter Zukerman (University of Washington)
- Piotr Przybyła (Universitat Pompeu Fabra)
- Samuel Attwood (Manchester Metropolitan University)
- Sergiu Nisioi (University of Bucharest)
- Xia Cui (Manchester Metropolitan University)

Table of Contents