Yuxia Wang | Resumé

MBZUAI NLP Postdoc, **57** papers with **21 first** (co-)author publications
Supervisors: Preslav Nakov, Timothy Baldwin, Karin Verspoor
✉ yuxia.wang@mbzuai.ac.ae

Education

The University of Melbourne — **Melbourne, Australia**
Doctor of Philosophy — *Sept, 2018–Jan, 2023*
Semantic Textual Similarity, Uncertainty Estimation

Beijing Institute of Technology — **Beijing, China**
Master of Computer Science and Technology — *Sept, 2016–July, 2018*

Beijing Institute of Technology — **Beijing, China**
Bachelor of Engineering in Software Engineering — *Sept, 2012–July, 2016*

Beijing Institute of Technology — **Beijing, China**
Bachelor of Economics in Economics (minor-degree) — *Sept, 2014–July, 2016*

Karlsruher Institute of Technology — **Karlsruher, Germany**
Exchange Student — *March, 2016–June, 2016*

Selected Publications

Yuxia Wang, Minghan Wang, Hasan Iqbal, Georgi Georgiev, Jiahui Geng, Iryna Gurevych, Preslav Nakov. OpenFactCheck: Building, Benchmarking Customized Fact-Checking Systems and Evaluating the Factuality of Claims and LLMs. COLING 2025.

Muhammad Arslan Manzoor*, **Yuxia Wang***, Minghan Wang, Preslav Nakov. Can Machines Resonate with Humans? Evaluating the Emotional and Empathic Comprehension of LMs. EMNLP 2024 findings.

Minghan Wang, **Yuxia Wang**, Thuy-Trang Vu, Ehsan Shareghi, Gholamreza Haffari. Exploring the Potential of Multimodal LLM with Knowledge-Intensive Multimodal ASR. EMNLP 2024 findings.

Hasan Iqbal*, **Yuxia Wang***, Minghan Wang, Georgi Georgiev, Jiahui Geng, Preslav Nakov. OpenFactCheck: A Unified Framework for Factuality Evaluation of LLMs. EMNLP 2024 Demo.

Yuxia Wang, Minghan Wang, Muhammad Arslan Manzoor, Georgi Georgiev, Rocktim Jyoti Das, Preslav Nakov. Factuality of large language models in the year 2024. EMNLP 2024 main.

Yuxia Wang, Revanth Gangi Reddy, Zain Muhammad Mujahid, Arnav Arora, Aleksandr Rubashevskii, Jiahui Geng, Osama Mohammed Afzal, Nadav Borenstein, Aditya Pillai, Isabelle Augenstein, Iryna Gurevych, Preslav Nakov. Factcheck-Bench: Fine-Grained Evaluation Benchmark for Automatic Fact-checkers. EMNLP 2024 findings.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, Thomas Arnold, Alham Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. M4GT-Bench: Evaluation Benchmark for Black-Box

Machine-Generated Text Detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL Volume 1: Long Papers)*, pages 3964–3992, Bangkok, Thailand. Association for Computational Linguistics. 2024.

Yuxia Wang, Zenan Zhai, Haonan Li, Xudong Han, Shom Lin, Zhenxuan Zhang, Angela Zhao, Preslav Nakov, and Timothy Baldwin. A Chinese Dataset for Evaluating the Safeguards in Large Language Models. In *Findings of the Association for Computational Linguistics (ACL)*, pages 3964–3992, Bangkok, Thailand. Association for Computational Linguistics. 2024.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, Thomas Arnold, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, Preslav Nakov. SemEval-2024 Task 8: Multidomain, Multimodel and Multilingual Machine-Generated Text Detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2057–2079, Mexico City, Mexico. Association for Computational Linguistics. 2024.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Alham Fikri Aji, and Preslav Nakov. M4: Multi-generator, Multi-domain, and Multi-lingual Black-Box Machine-Generated Text Detection. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics* ***(EACL Best Resource Paper Award)***, pages 1369–1407, St. Julian's, Malta. Association for Computational Linguistics, 2024.

Yuxia Wang, Haonan Li, Xudong Han, Preslav Nakov, and Timothy Baldwin. Do-Not-Answer: Evaluating Safeguards in LLMs. In *Findings of the Association for Computational Linguistics (EACL)*, pages 896–911, St. Julian's, Malta. Association for Computational Linguistics, 2024.

Yuxia Wang, Minghan Wang, and Preslav Nakov. Rethinking STS and NLI in Large Language Models. In *Findings of the Association for Computational Linguistics (EACL)*, pages 965–982, St. Julian's, Malta. Association for Computational Linguistics, 2024.

Yuxia Wang, Shimin Tao, Ning Xie, Hao Yang, Timothy Baldwin, and Karin Verspoor. Uncertainty-aware Semantic Textual Similarity. In *Transactions of the Association for Computational Linguistics (TACL)*, vol 11, pages 997–1013, Online, 2023.

Yuxia Wang, Daniel Beck, Timothy Baldwin, and Karin Verspoor. Uncertainty Estimation and Reduction of Pre-trained Models for Text Regression. In *Transactions of the Association for Computational Linguistics (TACL)*, vol 10, pages 680–696, Online, 2022.

Yuxia Wang, Minghan Wang, Yimeng Chen, Shimin Tao, Jiaxin Guo, Chang Su, Min Zhang and Hao Yang. Capture human disagreement distributions by calibrated networks for natural language inference. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022)*, pages 1524–1535, Dublin, Ireland, 2022.

Yuxia Wang, Timothy Baldwin, and Karin Verspoor. Noisy Label Regularisation for Textual Regression. In *The 29th International Conference on Computational Linguistics (COLING 2022)*, pages 4228–4240, Gyeongju, Republic of Korea, 2022.

Significant Research Contributions

Machine-generated Text Detection

Three Shared Tasks, Five Papers (One Best Resource Paper Award), 30 Collaborators

To safeguard human from the misuse of LLMs, we aim to advance more robust and generalized machine-generated text (MGT) detection methods. Since 2023, I have led two **shared tasks**: SemEval 2024 Task 8 and COLING 2025 GenAI Content Detection Task 1, and we are co-organizing CLEF 2025 PAN Lab Subtask 2. We have released three large-scale **datasets** designed for multilingual, multi-domain, and multi-generator MGT detection, alongside publishing **five papers** in ACL, EACL, EMNLP, and COLING. Additionally, a study examining human ability to identify AI outputs is set to be submitted to ACL 2025 in February. These initiatives are expected to enhance the generalization and robustness of MGT detection systems.

Multilingual LLM Safeguard Evaluation and Alignment

Six papers, Do-not-answer dataset in Stanford AI Index Report 2024, 8 languages

In August 2023, we developed **Do-not-Answer**, the first open-sourced LLM safety evaluation dataset, which was featured in the 2024 Stanford AI Index Report. Beyond straightforward questions where direct answers would be harmful, we include evasively framed risky questions to assess LLMs' risk perception, harmless questions with sensitive words to eight languages including English, Chinese, Arabic. Hindi, Russian, Kazakh, Bulgarian, and German, significantly contributing to the evaluation and improvement of LLM safety mechanisms.

LLM Factuality Evaluation and Improvement

Six Papers involving Dataset, Survey, Case Study, Method, and Demos

I explored LLM factuality from four perspectives and published six papers.

- Identify and rectify LLM factual errors by by model self-correction and low-latency evidence retrieval
- Flexibly customize automatic fact-checking pipelines according to application context
- Fine-grained evaluation of automatic fact-checking system performance step by step
- Easy access for general user to detect factual errors of model outputs (Demos)

Experience

MBZUAI — **Abu Dhabi, UAE**

Research Fellow (postdoc) — *Jan, 2023–current*

Research Topic:

- LLM factuality, safety and empathy evaluation and enhancement
- Machine-generated text detection
- Low-resource LLM development, particularly safety alignment and cultural bias
- Multimodal sarcasm recognition and empathetic response generation

LibrAI Startup — **Abu Dhabi, UAE**

Co-founder (COO) — *April, 2023–April, 2024*

Research Topic: Fact-checking and LLM-safety products.

Huawei 2012 Lab — **Beijing, China**

Research Intern — *May, 2021–May, 2022*

Research Topic:

- Natural language inference (NLI) and sentence semantic embedding
- Machine translation (MT), speech translation and quality evaluation

Awards and Achievements

- EACL 2024 Best Resource Paper Award, 2024
- CSC-University of Melbourne Research Scholarship, 2018
- National Scholarship, three times, BIT, China, 2012-2016

Languages

Name	Proficiency
Chinese	Native Proficiency
English	Full Professional Proficiency

Misc

- Primary language: Python, C
- Knowledge of: PyTorch, LaTeX, Linux
- Google Scholar: `http://bit.ly/3WqamfH`
- GitHub: `https://github.com/yuxiaw/`