MINGHAO WU

+61 452 231 543
minghao.wu@monash.edu
https://minghao-wu.github.io/

HIGHLIGHTS

- 8+ years of experience in research and development of machine learning, natural language processing (NLP), and generative AI.
- Strong publication record in top-tier conferences, such as ACL, EMNLP, COLING, and EACL.
- Programming Languages: Python, Java, and R.

EDUCATION

Monash University *Dec. 2021 - Jul. 2025 (expected)*
Doctor of Philosophy in Computer Science
Advisors: Gholamreza Haffari, George Foster, Lizhen Qu, and Trang Vu
Research Interests: Deep Learning, Natural Language Processing, Large Language Models, Language Agents, Multilinguality, Machine Translation

The University of Melbourne *Mar. 2016 - Jul. 2018*
Master of Information Technology in Computing
Advisors: Trevor Cohn

The University of Sydney *Mar. 2013 - Mar. 2016*
Bachelor of Science in Information Systems

EXPERIENCE

Research Intern *Nov. 2024 - Apr. 2025 (expected)*
Alibaba Group, China

- Worked on a project aimed at constructing a comprehensive multilingual benchmark covering more than 200 languages. This project involved collecting existing multilingual benchmarks, translating English benchmarks to other languages, and evaluating recent LLMs on the new benchmark.

Research Intern *Jul. 2023 - Oct. 2023*
Tencent AI Lab, China

- Worked on a project aimed at enhancing the capabilities of existing English-centric large language models (LLMs) by extending their linguistic coverage to include 150 natural languages and 150 programming languages. This involved further pretraining of recently released open-source LLMs on a vast collection of text and code corpora. The upgraded LLMs demonstrated the state-of-the-art performance across various multilingual evaluation benchmarks.

Visiting Researcher *Apr. 2023 - Jul. 2023*
Mohamed bin Zayed University of Artificial Intelligence, UAE

- Contributed to three research projects: (1) Investigated methods for the strategic compression of large generative models, successfully achieving significant reductions in model size without compromising their effectiveness; (2) Carried out a thorough assessment of biases present in large language models (LLMs) and human evaluators in judging machine-generated text; (3) Developed efficient techniques for distilling multilingual knowledge from large generative models into more compact versions.

Research Intern *Jul. 2020 - Jul. 2021*
Huawei Noah's Ark Lab, China

- Participated in two projects: (1) Implemented dynamic balancing techniques for the distribution of multiple datasets to optimize the training of multilingual and multi-domain machine translation systems; (2) Focused on pretraining both autoregressive and non-autoregressive multilingual machine translation systems using extensive parallel corpora.

Research Engineer *Aug. 2018 - Aug. 2019*
JD AI Research, China

- Developed the initial version of a conversational AI for an online shopping system, which involved creating an intent classification model, a coarse-grained answer search engine, and a fine-grained ranking model.

SELECTED PUBLICATIONS

- **Minghao Wu**, Thuy-Trang Vu, Lizhen Qu, and Gholamreza Haffari. "*The Best of Both Worlds: Bridging Quality and Diversity in Data Selection with Bipartite Graph* ." 2024.
- **Minghao Wu**, and Alham Fikri Aji. "*Style Over Substance: Evaluation Biases for Large Language Models*." In Proceedings of the 31th International Conference on Computational Linguistics (COLING). 2025. International Committee on Computational Linguistics.
- **Minghao Wu**, Thuy-Trang Vu, Lizhen Qu, and Gholamreza Haffari. "*Mixture-of-Skills: Learning to Optimize Data Usage for Fine-Tuning Large Language Models*." In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2024. Association for Computational Linguistics.
- **Minghao Wu**, Yulin Yuan, Gholamreza Haffari, and Longyue Wang. "*(Perhaps) Beyond Human Translation: Harnessing Multi-Agent Collaboration for Translating Ultra-Long Literary Texts*." 2024.
- **Minghao Wu**, Abdul Waheed, Chiyu Zhang, Muhammad Abdul-Mageed, and Alham Fikri Aji. "*LaMini-LM: A Diverse Herd of Distilled Models from Large-Scale Instructions*." In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL). 2024. Association for Computational Linguistics.
- **Minghao Wu**, Yufei Wang, George Foster, Lizhen Qu, and Gholamreza Haffari. "*Importance-Aware Data Augmentation for Document-Level Neural Machine Translation*." In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL). 2024. Association for Computational Linguistics.
- **Minghao Wu**, George Foster, Lizhen Qu, and Gholamreza Haffari. "*Document Flattening: Beyond Concatenating Context for Document-Level Neural Machine Translation*." In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL). 2023. Association for Computational Linguistics.
- **Minghao Wu**, Yitong Li, Meng Zhang, Liangyou Li, Gholamreza Haffari, and Qun Liu. "*Uncertainty-Aware Balancing for Multilingual and Multi-Domain Neural Machine Translation Training*." In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2021. Association for Computational Linguistics.
- **Minghao Wu**, Fei Liu, and Trevor Cohn. "*Evaluating the Utility of Hand-crafted Features in Sequence Labelling*." In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2018. Association for Computational Linguistics.