Dr. Kai Hui
Staff Research Scientist
Google DeepMind
Email:kai.hui.bj@gmail.com
Homepage|Google Scholar|DBLP|Semantic Scholar

(a) Personal Profile

I am a Staff Research Scientist at Google DeepMind, specializing in the pre-training and post-training of foundation models. My research advances LLMs through innovative data strategies and novel agentic designs. These technical contributions have supported the training of the Gemini model family and deep research products. Beyond my applied work, I have co-authored over 30 peer-reviewed papers and serve the academic community as a program committee member, editorial board member, and reviewer.

(b) Education & Training

Saarland University, Saarbrücken, Germany	Ph.D. in Computer Science
University of Chinese Academy of Sciences, Beijing, China	M.Sc. in Computer Science
Beijing Jiaotong University, Beijing, China	B.Sc. in Management Science

(c) Experiences

2024 – present	Research Scientist, Google DeepMind
2021 – 2024	Research Software Engineer, Google Research
2019 – 2021	Machine Learning Scientist, Amazon Alexa AI Search
2017 – 2019	Data Scientist, Cluster of Excellence for Deep Learning in SAP SE
2013 – 2017	Doctoral Researcher, Max Planck Institute for Informatics

(d) Selected Projects

1. Data-Centric Pre-training: Innovated data annotation strategies deployed in the Gemini model family to optimize performance and knowledge.
2. Inference-Time Scaling: Developed backtracking methods to solve multi-step reasoning challenges, successfully integrated into deep research products.

(e) Selected Publications

1. H. Zeng, **K. Hui**, H. Zhuang, Z. Qin, Z. Yue, H. Zamani, and D. Alon, "Can pre-training indicators reliably predict fine-tuning outcomes of llms?," in *arXiv preprint*, 2025.
2. Z. Qin, R. Jagerman, **K. Hui**, H. Zhuang, J. Wu, L. Yan, J. Shen, T. Liu, J. Liu, D. Metzler, X. Wang, and M. Bendersky, "Large language models are effective text rankers with pairwise ranking prompting," in *Findings of NAACL 2024*, ACL.
3. J. Lee, Z. Dai, X. Ren, B. Chen, D. Cer, J. R. Cole, **K. Hui**, M. Boratko, R. Kapadia, W. Ding, Y. Luan, S. M. K. Duddu, G. H. Abrego, W. Shi, N. Gupta, A. Kusupati, P. Jain, S. R. Jonnalagadda, M.-W. Chang, and I. Naim, "Gecko: Versatile text embeddings distilled from large language models," in *arXiv preprint*, 2024.
4. R. Pradeep, **K. Hui**, J. Gupta, A. Lelkes, H. Zhuang, J. Lin, D. Metzler, and V. Tran, "How does generative retrieval scale to millions of passages?," in *EMNLP 2023*, ACL.
5. **K. Hui**, A. Yates, K. Berberich, and G. de Melo, "PACRR: A position-aware neural ir model for relevance matching," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, EMNLP 2017, ACL.

I published more than 30 papers. Refer to my Google Scholar page for the full list.