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

(a) Personal Profile

I work on pre-training, post-training, and agent systems for large language models, with a focus on developing novel methods to improve model capability and robustness. My recent work explores reinforcement learning and inference-time reasoning to improve adaptability and multi-step reasoning. My contributions include: (1) data-centric pre-training methods applied in Gemini model development; (2) novel methods for agent systems focusing on failure recovery and iterative reasoning in deep research settings; (3) reinforcement learning to mitigate rigid behaviors.

(b) Research Interests

- Reinforcement learning and evaluation for model behavior
- Context retrieval and selection in agent systems
- Scaling inference-time compute for long-horizon reasoning tasks

(c) Selected Projects

1. **Inference-Time Reasoning & Scaling.** Proposed backtracking methods for multi-step reasoning; improved performance in deep research settings and integrated into production agents.
2. **Data-Centric Pre-training.** Led metadata-driven data annotation; applied in Gemini smaller-tier models, improving downstream performance.
3. **SFT and RL for Model Behavior.** Exploring reinforcement learning methods to improve adaptability and reduce rigid behaviors; developing failure-aware training signals for robust model behavior.

(d) Experiences

2021 – present	Research Scientist, Google Research & Google DeepMind
2019 – 2021	Machine Learning Scientist, Amazon Alexa AI
2017 – 2019	Data Scientist, Cluster of Excellence for Deep Learning in SAP SE

(e) 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

(f) Selected Publications

1. H. Zeng, **K. Hui**, H. Zhuang, Z. Qin, Z. Yue, H. Zamani, and D. Alon, "LLM development at fixed scale: Lightweight proxies to overcome the perplexity illusion," 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.

See my **Google Scholar page** for the full list.