Kuang Wang

kuangwang@link.cuhk.edu.cn | wangkevin02 | Google Scholar | +86 150-8828-0634

Research Interests

Speech Language Models · Personalized Large Language Models · User Simulation · Conversational AI

Education

The Chinese University of Hong Kong, Shenzhen (CUHK-SZ) *Shenzhen, China*
Ph.D. in Computer Science *Advisor: Prof. Haizhou Li* *Sep. 2024 – Present*

Zhejiang University (ZJU) *Hangzhou, China*
B.Eng. in Bio-system Engineering *Sep. 2020 – Jun. 2024*

- **GPA:** 3.96/4.0 (89.98/100) **Rank:** 4/64
- **Honors & Awards:**
  - Outstanding Graduate, Zhejiang University *2024*
  - First-Class Scholarship(top 5%), Zhejiang University *2021*

Publications

Conference Papers

Wang, K., Li, X., Yang, S., et al.. **"Know You First and Be You Better: Modeling Human-Like User Simulators via Implicit Profiles."** *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025 **[ACL 2025]** PDF Code

Ke, R., Xu, J., Yang, S., **Wang, K.**, Jiang, F., Li, H.. **"CATCH: A Controllable Theme Detection Framework with Contextualized Clustering and Hierarchical Generation."** *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026 **[AAAI 2026]** PDF

Preprints

Jiang, F., **Wang, K.**, Li, H.. **"Bridging Research and Readers: A Multi-Modal Automated Academic Papers Interpretation System."** *arXiv preprint arXiv:2401.09150*, 2024 **[arXiv]** PDF Code Demo

Liu, S., Xu, J., Jiang, F., **Wang, K.**, Zhao, Z., Huang, C., Gu, J., Yin, C., Li, H.. **"Discourse-Aware Dual-Track Streaming Response for Low-Latency Spoken Dialogue Systems."** *arXiv preprint arXiv:2602.23266*, 2026 **[arXiv]** PDF

Internship Experience

Tencent — Multimodal Technology Center *Shenzhen, China*
Research Intern, CSIG · Yuanbao Product Dept. · Speech Language Model Group *Jul. 2025 – Present*

- **Objective:** Built a post-pretraining text data pipeline for Speech Language Model (SLM) pre-training, targeting preservation of semantic intelligence inherited from the base LLM during large-scale speech–text alignment pre-training
- **Work:** Designed and implemented a distributed, multi-node parallel data preparation pipeline based on **Ray**, covering the full workflow from raw corpora to training-ready data at hundred-terabyte scale, including language filtering, cleaning & normalization, quality scoring (perplexity- and classifier-based), domain classification & stratified sampling, and MinHash LSH deduplication
- **Impact:** Processed **100+ TB** of raw text; curated corpus used for continual pre-training of a text-based LLM into a **text–speech interleaved generation SLM**, powering the *Tencent Meeting* voice assistant

Teaching Experience

Teaching Assistant, *Reinforcement Learning* *Spring – Summer 2026*
School of Data Science, The Chinese University of Hong Kong, Shenzhen

Teaching Assistant, *Machine Learning* *Fall – Winter 2025*
School of Data Science, The Chinese University of Hong Kong, Shenzhen

Teaching Assistant, *Linear Algebra and Its Applications* *Spring – Summer 2025*
School of Data Science, The Chinese University of Hong Kong, Shenzhen

Skills

Programming: Python, C, PyTorch, Scikit-learn

Tools & Platforms: Git, LaTeX, Ray, Alibaba Cloud MaxCompute, Linux

Languages: Mandarin Chinese (Native), English (Fluent)