Kaili Huang

🎓 Google Scholar 💼 LinkedIn 🐙 GitHub

Email : kaili.khuang@gmail.com

Mobile : +1-650-665-4877

EDUCATION

Stanford University — Stanford, CA
Master of Science in Computer Science; GPA: 4.0 — *Sep. 2021 – Jun. 2023*

Tsinghua University — Beijing, China
Bachelor of Engineering in Industrial Engineering; CS GPA: 3.8 — *Aug. 2016 – Jul. 2020*

EXPERIENCE

Microsoft — Redmond, WA
Applied Scientist — *Aug. 2023 - Present, 2 years*

- **Next-Gen Multi-Modal Retrieval for Ads**: Led research on CLIP, SigLIP, and Perception Encoder across five multi-modal product ads tasks involving both text and image inputs. Independently built the full post-training pipeline, including data preprocessing, custom dataloaders, model adaptation (based on CLIP), training, and evaluation, to enable extensive experimentation. Trained the first teacher model and distilled a lightweight student, achieving 28–34% AUC gains. The student model reduced defect rate by up to 6.2% and was mainstreamed to improve retrieval across multiple downstream ads systems.

- **LLM-Powered Retrieval-Augmented Generation in Ads Copilot**: Designed and deployed LLM-based evaluation and retrieval improvements for Ads Copilot Q&A product, driving 15.6% Recall@10 improvement and 6% overall answer quality gain. Established prompt-based Q&A evaluation pipelines (F1=0.94 for plugin accuracy and 0.89 for answer correctness), developed human-reviewed eval sets, and integrated core metrics into production portals. Contributed to product launches including multi-round conversation support, user intent rewriting feature and user frustration handling function.

Microsoft — Redmond, WA
Data Scientist Intern — *Jun. 2022 - Sep. 2022, 3 months*

- **Highly Efficient Multi-Modal Transformers via Structured Pruning**: Reduced CLIP model size by 40% with minimal accuracy loss (-1%) by extending structured pruning to multi-modal settings. Built sparse training pipeline, decomposed Block Movement Pruning into modular steps, and introduced auxiliary losses to stabilize training and balance modalities.

ByteDance (TikTok) — Beijing, China
Machine Learning Engineer — *Jul. 2020 - Aug. 2021, 1 year*

- **End-to-End Fake News Detection System**: Developed an end-to-end system detecting around 100 fake news articles daily, optimized for high recall to support downstream human verification. Tuned for high recall to ensure coverage of harmful content, achieving 50% precision with downstream human review safeguarding false positives. Collected 2M high-quality human labels and trained BERT classification and NLI models, achieving F1 up to 0.72. Innovatively applied symbolic learning for numeric reasoning (+15% F1). Deployed models into production with auto-update and RPC integration.

- **Bot-Written Articles Detection**: Synthesized articles with GPT models and built BERT-based models to identify AI-generated articles, achieving an F1 score of 0.98.

PUBLICATIONS

- **Kaili Huang** et al. "ColBERT-serve: Efficient Multi-Stage Memory-Mapped Scoring." *ECIR 2025.* 🐙 📄
- C Gunasekara, et al. "Overview of the ninth dialog system technology challenge: Dstc9." *IEEE/ACM Transactions on Audio, Speech, and Language Processing 2024.* 📄
- Jinchao Li, Qi Zhu, Lingxiao Luo, Lars Liden, **Kaili Huang**, etc. "Multi-Domain Task Completion Dialog Challenge II at DSTC9. " *AAAI 2021.* 📄
- Qi Zhu, **Kaili Huang**, Zheng Zhang, Xiaoyan Zhu, and Minlie Huang. "CrossWOZ: A Large-Scale Chinese Cross-Domain Task-Oriented Dialogue Dataset." *TACL 2020.* 🐙 📄
- Hao Zhou, Chujie Zheng, **Kaili Huang**, Minlie Huang, and Xiaoyan Zhu. "KdConv: A Chinese Multi-domain Dialogue Dataset Towards Multi-turn Knowledge-driven Conversation." *ACL 2020.* 🐙 📄
- Yida Wang, Pei Ke, Yinhe Zheng, **Kaili Huang**, Yong Jiang, Xiaoyan Zhu, and Minlie Huang. "A Large-Scale Chinese Short-Text Conversation Dataset." *NLPCC 2020 (Best Student Paper Award).* 🐙 📄

Teaching & Academic Services

- **Stanford CS224n Natural Language Processing with Deep Learning**: Mentored 10+ groups of students for the final projects; held weekly office hours; wrote lecture notes; designed and graded assignments. 🔗

- **Paper Review**: Conducted 30+ paper review work for top-tier conferences and journals in natural language processing (NLP) and computer vision (CV), including EMNLP'23, ICLR'24, WACV'24, COLING'24, SIGIR'24, KDD"24, CIKM'24, WACV'25, ICLR'25, COLING'25, ACL'25, Computer Speech & Language.

Projects

- **Engineering Effective In-Context Inputs for GPT-3 in OpenQA**: Designed and evaluated novel in-context learning strategies to improve GPT-3's performance on OpenQA without access to gold passages. Explored lexical, syntactic, and semantic similarity-based example selection methods, and introduced reverse ordering to enhance contextual relevance. The semantic similarity + reverse order strategy achieved the best performance (F1: 0.57), yielding a 5% improvement over the random baseline. Findings highlight the impact of example amount, quality, similarity, and ordering on large language model effectiveness. 📄

- **Optimizing Dialogue History Encoding for Multilingual Task-Oriented Systems**: Proposed an efficient training strategy for multilingual virtual assistants by replacing natural language dialogue history with structured dialogue states. Built on the BiToD architecture to reduce reliance on weak natural language encoders and improve slot value extraction. The work investigated the impact of history length (number of previous turns) on model performance, identified diminishing returns, and explored few-shot learning for hard examples. It is designed to improve scalability and robustness in low-resource languages. 📄

Awards

- **NLPCC Best Student Paper** *2020*
- **Stanford UGVR Scholar (Up to 18 students from China are admitted per year)** *2019*
- **1st Prize in National Olympiad in Informatics in Provinces** *2014*

Skills

- **Languages**: Scala, Python, Javascript, C++, SQL, Java

- **Frameworks & Tools**: PyTorch, TensorFlow, Hadoop, Hive, Spark, GCP, Azure, MongoDB, Django, Git