Kaili Huang

Google Scholar (cited 490+) LinkedIn GitHub (starred 3K+)

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**: Led research on CLIP, SigLIP, and Perception Encoder across five real-world multi-modal advertising scenarios involving text-to-image, image-to-image, and text-to-text retrieval. Built multi-modal models entirely from scratch, including custom CLIP-based architectures, data preprocessing pipelines, distributed training code, evaluation scripts, and configuration management. Trained the first teacher model and finetuned a lightweight student using distillation, achieving 28–34% AUC gains compared to the baseline (production model). The student model reduced defect rate by up to 6.2% in online experiments and was successfully deployed to production after A/B testing.
  - **LLM-Powered Retrieval-Augmented Generation in Ads Copilot**: Designed and deployed LLM retrieval-augmented generation (RAG) systems with prompt-based reward proxy evaluation, chain-of-thought prompting, and human-in-the-loop metrics. Developed modular rubric-based evaluation protocols to benchmark RAG performance and guard against failure cases (e.g., hallucination, frustration). Curated high-quality, human-reviewed evaluation sets and improved the retrieval algorithm, achieving a 15.6% Recall@10 increase and a 6% overall answer quality gain. Contributed to production-level LLM features, including multi-turn conversation support, user intent rewriting, and frustration detection, and integrated core metrics into real-time dashboards.
- **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. The model was tuned for high recall to ensure coverage of harmful content, achieving 50% precision with downstream human review that safeguarded against false positives. Led the collection of 2M high-quality human annotations by collaborating with in-house labeling teams via a large-scale data creation platform. Designed and iteratively refined labeling guidelines, coordinated daily and weekly reviews with annotation leads, and ensured consistent data quality and task coverage at scale. Trained BERT classification and NLI models, and achieved F1 up to 0.72. Innovatively applied symbolic learning for numeric reasoning (+15% F1). Deployed models into production with auto-retraining, version-controlled rollout, and RPC integration.
- **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 a sparse training pipeline, decomposed Block Movement Pruning into modular steps, and introduced auxiliary losses to stabilize training and balance modalities.

Selected Publications

- **Kaili Huang**, Thejas Venkatesh, Uma Dingankar, Antonio Mallia, Daniel Campos, Jian Jiao, Christopher Potts, Matei Zaharia, Kwabena Boahen, Omar Khattab, Saarthak Sarup, Keshav Santhanam. "ColBERT-serve: Efficient Multi-Stage Memory-Mapped Scoring." *ECIR 2025.*
- 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**).*

Research Projects

- **Task-Oriented Dialogue Systems via Reinforcement Learning (RL)**:
  - Research Assistant. Advisor: Tengyu Ma, Stanford University
  - Stanford Undergraduate Visiting Research Program (UGVR) Program
  - Used the RL algorithm Stochastic Lower Bound Optimization (SLBO), and built an agent for the movie-ticket booking task. Built a Vanilla Policy Gradient (VPG) agent and created a chatbot environment to train the agent. Built a pipeline for taking user simulator samples and training dialogue policies on different model-based deep reinforcement learning (RL) algorithms.

- **Optimizing Dialogue History Encoding for Multilingual Virtual Assistant**: Optimized multilingual task-oriented dialogue agents, improving low-resource scalability and system interpretability which is key to virtual assistant deployment in enterprise settings. 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. 📄

- **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. 📄

Teaching & Academic Services

- **Stanford CS224n Natural Language Processing with Deep Learning**:
  - Instructor: Christopher Manning.
  - 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.

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

- **Programming Languages**: Python (primary, production-ready), C++, Java, R, SQL, JavaScript
- **ML & LLM Tooling**: PyTorch, TensorFlow, scikit-learn, Pandas, NumPy, LangChain, RL, RAG, Prompt Engineering, Evaluation Frameworks
- **APIs & Infrastructure**: OpenAI APIs, Enterprise API Integration, Spark, Hadoop, Hive, MongoDB, Django
- **Cloud & DevOps**: GCP, Azure, Git, Docker