Hao Bai

MS in Computer Science, UIUC *(Advisor: Heng Ji)*

✉ haob2@illinois.edu • ⌂ jackgethome.com

Interests: Foundational Models, Representation Learning, Information Retrieval

Education

MS in Computer Science	University of Illinois, Urbana Champaign	Aug 2023 - May 2025
BE in Computer Engineering (Dual)	Zhejiang University, China	Sep 2019 - Jul 2023
BS in Computer Engineering (Dual)	University of Illinois, Urbana Champaign	Aug 2019 - May 2023

Professional Experience

University of Illinois at Urbana-Champaign   **May 2022 - Present**
Graduate Research Assistant, Blender Lab   *Champaign, US*
○ Working on projects in the areas of explainable language generation and large-scale information retrieval.

Microsoft Research   **Nov 2022 - May 2023**
Research Intern, Data, Knowledge, Information (DKI) Group   *Beijing, CN*
○ Worked on projects in the areas of incident root cause category prediction and root cause generation, with a special focus on LLM-based approaches.

Selected Papers (* denotes individual author)

White-Box Transformers via Sparse Rate Reduction: Compression Is All There Is? [PDF]   *JMLR Submission*
*Yaodong Yu, Sam Buchanan, Druv Pai, **Hao Bai**, Yuexiang Zhai, Yi Ma, et al.*   *UC Berkeley*
○ Proposed that a natural objective of representation learning is to compress and transform the distribution of the data towards a mixture of low-dimensional Gaussian distributions supported on incoherent subspaces, and empirically proved its correcness by proposing a new foundational model that achieves comparable performance across various domains.

Progressive Responses with Real-Time Internet Search for Conversations [PDF]   *WSDM'24*
*Revanth Reddy, Sharath Chandra, **Hao Bai**, Wentao Yao, Chengxiang Zhai, et al.*   *Amazon Alexa Grant*
○ Introduces the use of progressive response generation to effortlessly blend search results into the bot's responses, while ensuring low response latency, which cuts down user waiting times by 50%.

Social Conversational Commonsense-Guided Search Query Generation [PDF]   *EMNLP'23*
*Revanth Reddy, **Hao Bai**, Wentao Yao, Sharath Chandra, Heng Ji, ChengXiang Zhai*   *Amazon Alexa Grant*
○ Proposed to integrate **social commonsense knowledge** to the **query generator** by first generating initial responses from a commonsense response generator, followed by distilling knowledge from LLM, which achieves state-of-the-art performance on the quality of the generated query and also downstream tasks like final response.

MedoFlow: An Educational Software Framework for Deep Learning*   *Thesis*
○ Implemented a complete software framework for deep learning **from scratch** using Python frontend and TVM backend based on TinyFlow, which achieves the **same accuracy** and a **comparable time efficiency** with state-of-the-art frameworks like PyTorch and TensorFlow on fundamental applications like MLP, CNN and RNN.