WENDA XU

Research Scientist Specializing in Large Language Model Evaluation and Post-training

wendaxu@ucsb.edu +1-972-343-8224 Mountain View, CA https://xu1998hz.github.io/
https://github.com/xu1998hz/ WendaXu2 https://www.linkedin.com/in/wenda-xu-866040163/

EDUCATION

Ph.D., Computer Science (GPA: 3.8/4.0)

University of California, Santa Barbara

Sept 2020 – Mar 2025 Santa Barbara, CA

- Disseration: On Evaluation and Efficient Post-training for LLMs
- Advisors: William Yang Wang, Ph.D., Lei Li, Ph.D.

B.S., Computer Science (GPA: 3.9/4.0)

University of California, Davis

Sept 2016 – Mar 2020 Davis, CA

- Senior Design Project: Visual SLAM using ORB-SLAM2 with Path Finding **(Best Senior Design of a Year)**
- Advisors: Chen-Nee Chuah, Ph.D.

RESEARCH INTERESTS

My research interests lie in the area of large language model (LLM) evaluation and efficient post-training. I am the first author of SEScore1&2 and InstructScore **(Best Unsupervised Text Generation metrics at WMT22 shared task)**. Currently, I am actively working on LLM post-training techniques, in both preference learning and knowledge distillation.

FIRST AUTHOR PUBLICATIONS

Wenda Xu, Rujun Han, Zifeng Wang, Long Le, Dhruv Madeka, Lei Li, William Yang Wang, Rishabh Agarwal, Chen-Yu Lee, Tomas Pfister. *Speculative Knowledge Distillation: Bridging the Teacher-Student Gap Through Interleaved Sampling*. ICLR 2025. A generic KD framework that generalizes to on-policy and supervised KD, achieves substantial gains in task specific and task agnostic knowledge distillation.

Wenda Xu*, Jiachen Li*, William Yang Wang, Lei Li. *BPO: Supercharging Online Preference Learning by Adhering to the Proximity of Behavior LLM*. EMNLP 2024 (*equal contribution). On-policy BPO achieves superior performance compared to DPO on TL;DR (89.5% vs 72.0%), Helpfulness (93.5% vs 82.2%), and Harmfulness (97.7% vs 77.5%).

Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, William Yang Wang. *Pride and Prejudice: LLM Amplifies Self-Bias in Self-Refinement*. ACL 2024 Oral. The first study to define and quantify the bias exhibited by LLMs when assessing their own generated outputs.

Wenda Xu, Daniel Deutsch, Mara Finkelstein, Juraj Juraska, Biao Zhang, Zhongtao Liu, William Yang Wang, Lei Li, Markus Freitag. *LLMRefine: Pinpointing and Refining Large Language Models via Fine-Grained Actionable Feedback*. NAACL 2024. An inference time technique iteratively improves PALM2 for 1.7 MetricX on translation tasks, 8.1 ROUGE-L on ASQA and 2.2 ROUGE-L on topical summarization.

Wenda Xu, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Yang Wang, Lei Li. *INSTRUCTSCORE: Explainable Text Generation Evaluation with Finegrained Feedback*. EMNLP 2023 Oral. A fine-grained 7B LLM evaluator, trained on synthetic data, surpasses unsupervised metrics, including those based on 175B GPT-3 and GPT-4.

Wenda Xu, Xian Qian, Mingxuan Wang, Lei Li, William Yang Wang. *SEScore2: Learning Text Generation Evaluation via Synthesizing Realistic Mistakes*. ACL 2023. Kendall correlation improved 14.3% from SEScore.

Wenda Xu, Yilin Tuan, Yujie Lu, Michael Saxon, Lei Li, William Yang Wang. *Not All Errors are Equal: Learning Text Generation Metrics using Stratified Error Synthesis*. EMNLP 2022. SEScore: No.1 unsupervised metric at WMT22 metrics shared task.

Wenda Xu, Michael Saxon, Misha Sra, William Yang Wang. *Self-Supervised Knowledge Assimilation for Expert-Layman Text Style Transfer*. AAAI 2022. Relative improvement in overall success rate by 106%.

COLLABORATION PUBLICATIONS

Xi Xu, **Wenda Xu**, Siqi Ouyang, Lei Li. *CA*: Addressing Evaluation Pitfalls in Computation-Aware Latency for Simultaneous Speech Translation*. NAACL 2025

Juhyun Oh, Eunsu Kim, Jiseon Kim, **Wenda Xu**, Inha Cha, William Yang Wang, Alice Oh. *Uncovering Factor Level Preferences to Improve Human-Model Alignment*. On submission

Chinmay Dandekar, **Wenda Xu**, Xi Xu, Siqi Ouyang, Lei Li. *Translation Canvas: An Explainable Interface to Pinpoint and Analyze Translation Systems*. EMNLP Demo 2024

Liangming Pan, Michael Saxon, **Wenda Xu**, Deepak Nathani, Xinyi Wang, William Yang Wang. *Automatically Correcting Large Language Models: Surveying the landscape of diverse self-correction strategies*. TACL 2024

Michael Saxon, Xinyi Wang, **Wenda Xu**, William Yang Wang. *PECO: Examining Single Sentence Label Leakage in Natural Language Inference Datasets through Progressive Evaluation of Cluster Outliers*. EACL 2023

Yujie Lu, Weixi Feng, Wanrong Zhu, **Wenda Xu**, Xin Eric Wang, Miguel Eckstein, William Yang Wang. *Neuro-Symbolic Causal Language Planning with Commonsense Prompting*. ICLR 2023

Wanrong Zhu, An Yan, Yujie Lu, **Wenda Xu**, Xin Eric Wang, Miguel Eckstein, William Yang Wang. *Visualize Before You Write: Imagination-Guided Open-Ended Text Generation*. EACL 2023

Yi-Lin Tuan, Alon Albalak, **Wenda Xu**, Michael Saxon, Connor Pryor, Lise Getoor, William Yang Wang. *CausalDialogue: Modeling Utterance-level Causality in Conversations*. ACL 2023

INDUSTRY RESEARCH EXPERIENCE

Research Science Intern

Google Cloud Research

Jun 2024 – Oct 2024 | Los Angeles, CA

- Mentors: Rujun Han, Zifeng Wang, Rishabh Agarwal, Chen-Yu Lee.
- Publication: Speculative Knowledge Distillation: Bridging the Teacher-Student Gap Through Interleaved Sampling

Research Science Intern

Google Translate Research

Jun 2023 – Dec 2023 | Mountain View, CA

- Mentors: Dan Deutsch, Markus Freitag.
- Publication: LLMRefine: Pinpointing and Refining Large Language Models via Fine-Grained Actionable Feedback

Research Science Intern

TikTok AI Lab

Jun 2022 – Oct 2022 | Mountain View, CA

- Mentors: Xian Qian, Mingxuan Wang.
- Publication: SESCORE2: Learning Text Generation Evaluation via Synthesizing Realistic Mistakes

SKILLS

Software Proficiencies

Python | Pytorch | Numpy | C | C++ | Linux

Conceptual

Deep learning | Natural Language Processing | Large Language model (LLM) | LLM Evaluation | Post-training

HONORS

- The Robert Noyce Fellowship, Academic Excellence Fellowship (UC Santa Barbara, 2022)
- Honor Graduation (UC Davis, 2020)
- Thomas E. Bruzzone Award (UC Davis, 2019)
- Robert Murdoch Memorial Scholarship (UC Davis, 2019)
- Best Senior Design of a Year (Visual SLAM) (UC Davis, 2019)
- College of Engineering Dean's Honor list (UC Davis, 2016-2020)