

Wenda Xu

wendaxu@ucsb.edu

Education

University of California, Santa Barbara

Ph.D., Computer Science: 3.9/4.0

Advisor: William Yang Wang, Ph.D

Lei Li, Ph.D

Santa Barbara, CA

9/2020–3/2025

University of California, Davis

BS., Computer Science: 3.9/4.0

Senior Design Project—Visual SLAM using ORB-SLAM2 with Path Finding

Advisors: Chen-Nee Chuah, Ph.D.

Davis, CA

9/2016–3/2020

Research Interests

My major research interests lie in the area of text generation evaluation and large language model (LLM) alignment (at pre-training, post-training and inference stages). In one sentence, I want to learn metrics that can assess LLM's generation quality and align LLM with well defined feedback.

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's Publications & Preprints

1. **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”, <https://openreview.net/pdf?id=EgJhwYR2tB>, 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
2. **Wenda Xu***, Jiachen Li*, William Yang Wang, Lei Li, “BPO: Supercharging Online Preference Learning by Adhering to the Proximity of Behavior LLM”, <https://arxiv.org/abs/2406.12168>, EMNLP 2024, *equal contribution (**TL;DR (72.0%→89.5%), Helpfulness (82.2%→93.5%), Harmfulness (77.5%→97.7%)**)
3. **Wenda Xu**, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, William Yang Wang, “Pride and Prejudice: LLM Amplifies Self-Bias in Self-Refinement”, <https://arxiv.org/abs/2402.11436>, ACL 2024 Oral (**First define and quantify LLM's self-bias towards its own outputs**)
4. **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”, <https://arxiv.org/abs/2311.09336>, NAACL 2024 (**Fine-grained LLM agent iteratively improves PALM2 for 1.7 MetricX on translation tasks, 8.1 ROUGE-L on ASQA, 2.2 ROUGE-L on topical summarization**)
5. **Wenda Xu**, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Yang Wang, Lei Li, “INSTRUCTSCORE: Explainable Text Generation Evaluation with Finegrained Feedback”, <https://arxiv.org/abs/2305.14282>, EMNLP 2023 Oral (**Fine-grained 7B LLM evaluator surpasses all other unsupervised metrics, including those based on 175B GPT-3 and GPT-4**)
6. **Wenda Xu**, Xian Qian, Mingxuan Wang, Lei Li, William Yang Wang, “SEScore2: Learning Text Generation Evaluation via Synthesizing Realistic Mistakes”, <https://arxiv.org/abs/2212.09305>, ACL2023 (**The overall Kendall correlation improves 14.3% from SEScore**)
7. **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”, <https://arxiv.org/abs/2210.05035>, EMNLP 2022, **SEScore: No.1 metric among all unsupervised metrics in WMT22 metrics shared task**
8. **Wenda Xu**, Michael Saxon, Misha Sra and William Yang Wang, “Self-Supervised Knowledge Assimilation for Expert-Layman Text Style Transfer”, <https://arxiv.org/abs/2110.02950>, **relative improving overall success rate by 106%, AAAI 2022**

Collaboration Publications

9. Xi Xu, **Wenda Xu**, Siqi Ouyang, Lei Li, “CA*: Addressing Evaluation Pitfalls in Computation-Aware Latency for Simultaneous Speech Translation”, <https://arxiv.org/abs/2410.16011>, NAACL 2025
10. Juhyun Oh, Eunsu Kim, Jiseon Kim, **Wenda Xu**, Inha Cha, William Yang Wang, Alice Oh, “Uncovering Factor Level Preferences to Improve Human-Model Alignment”, <https://arxiv.org/pdf/2410.06965>, on submission
11. Chinmay Dandekar, **Wenda Xu**, Xi Xu, Siqi Ouyang, Lei Li, “Translation Canvas: An Explainable Interface to Pinpoint and Analyze Translation Systems”, <https://arxiv.org/abs/2410.10861>, EMNLP Demo 2024
12. 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”, <https://arxiv.org/pdf/2308.03188.pdf>, TACL 2024
13. 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”, <https://arxiv.org/abs/2112.09237>, EACL2023
14. Yujie Lu, Weixi Feng, Wanrong Zhu, **Wenda Xu**, Xin Eric Wang, Miguel Eckstein, William Yang Wang, “Neuro-Symbolic Causal Language Planning with Commonsense Prompting”, <https://arxiv.org/abs/2206.02928>, ICLR2023
15. 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”, <https://arxiv.org/pdf/2210.03765.pdf>, EACL2023
16. Yi-Lin Tuan, Alon Albalak, **Wenda Xu**, Michael Saxon, Connor Pryor, Lise Getoor, William Yang Wang, “CausalDialogue: Modeling Utterance-level Causality in Conversations”, <https://arxiv.org/pdf/2212.10515.pdf>, ACL2023

Industry Research Experience

Google Cloud Research

Los Angeles, CA

Research Science Intern

6/2024 - 10/2024

Mentors: Rujun Han, Zifeng Wang, Rishabh Agarwal, Chen-Yu Lee.

Publication: [Speculative Knowledge Distillation: Bridging the Teacher-Student Gap Through Interleaved Sampling](#)

Google Translate Research

Mountain View, CA

Research Science Intern

6/2023 - 12/2023

Mentors: Dan Deutsch, Markus Freitag.

Publication: [LLMRefine: Pinpointing and Refining Large Language Models via Fine-Grained Actionable Feedback](#)

TikTok AI Lab

Mountain View, CA

Research Science Intern

6/2022 - 10/2022

Mentors: Xian Qian, Mingxuan Wang.

Publication: [SESCORE2: Learning Text Generation Evaluation via Synthesizing Realistic Mistakes](#)

Skills

Software Proficiencies

Python (Pytorch, Tensorflow, Numpy, SciPy, Sklearn, Huggingface etc), C, C++, Linux, PHP

Conceptual

Deep learning, Natural Language Processing (NLP), Text Generation, Language model

Selected Coursework

Probability; Matrix Analysis; Machine Learning; Algorithm and Data Structure; Machine Translation; Natural Language Processing; Computer Vision; Computer Graphics; Self-Driving; Combinatorics and Graph Theory.

Honors

UCSB, The Robert Noyce Fellowship, Academic Excellence Fellowship	2022
UC Davis, Honor Graduation	2020
UC Davis, Thomas E. Bruzzone	2019
UC Davis, Robert Murdoch Memorial Scholarship	2019
UC Davis, Best Senior Design of a year (Visual SLAM)	2019
UC Davis, College of Engineering, Dean's Honor list	16-20

Last updated Jan 30, 2025