

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

9/2020–6/2025

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 (using data augmentation, training and inference time approaches). I want to learn metrics that can assess LLM's generation quality (both in the form of quality score or natural language diagnostic report) and align LLM with human principles.

I am the first author of SEScore1&2 and InstructScore (**Best Unsupervised metrics at WMT22 shared task**). Currently, I am actively working on improving training time approach to align LLM with well defined metric, reward or human feedback.

First Author's Publications & Preprints

1. **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%)**)
2. **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**)
3. **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**)
4. **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**)
5. **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**)
6. **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**
7. **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

8. 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

9. 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
10. 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
11. 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
12. 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 - Present

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

- Knowledge distillation of LLM - ongoing project

Google Translate Research

Mountain View, CA

Research Science Intern

6/2023 - 12/2023

Mentors: Markus Freitag, Dan Deutsch.

- Used a learned fine-grained feedback model (InstructScore style) to pinpoint defects. Using original LLM (PALM2) as a proposal of edits, **LLMRefine** searches for defect-less text via simulated annealing. LLMRefine improves 1.7 MetricX on translation tasks, 8.1 ROUGE-L on ASQA, 2.2 ROUGE-L on topical summarization [3].

TikTok AI Lab

Mountain View, CA

Research Science Intern

6/2022 - 10/2022

Mentors: Mingxuan Wang, Xian Qian.

- Synthesized realistic model mistakes by perturbing sentences retrieved from a corpus. Developed a self-supervised technique to train a learned metric to estimate number of errors and severity levels in each sample; **SEScore2(14.3% improvements from SCScore)** achieves top performance in Machine and Speech Translation and data-to-text [5]

Skills

Software Proficiencies

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

Conceptual

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

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 + 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