WENDA XU

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

wendax@google.com +1-972-343-8224 Mountain View, CA Webpage Github
Twitter Linkedin

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 focuses on improving large language models (LLMs) through rigorous evaluation and efficient post-training. I actively develop automated methods to assess model capabilities, including scalable and efficient data curation technique for building model-specific challenge benchmarks [1] [2] [3]. To complement this, I design unsupervised and explainable evaluation metrics. My work in this area led to SEScore1&2 [11][12] and InstructScore [9][10], which was recognized as the best unsupervised metric at the WMT22 shared task. My background also includes developing efficient post-training techniques, from preference learning with BPO [5] to knowledge distillation with Speculative KD [4].

SELECTED PUBLICATIONS

A full list of my publications is available on my Google Scholar profile.

[1] **Wenda Xu***, Vilém Zouhar*, Parker Riley, Mara Finkelstein, Markus Freitag, Daniel Deutsch. *Searching for Difficult-to-Translate Test Examples at Scale*. On submission. A scalable and efficient automatic data curation technique for building model-specific challenge benchmarks by finding the most difficult test examples.

[2] **Wenda Xu**, Sweta Agrawal, Vilém Zouhar, Markus Freitag, Daniel Deutsch. *Deconstructing Self-Bias in LLM-generated Translation Benchmarks*. On submission. We demonstrate that LLM-generated benchmarks exhibit a systematic self-bias that inflates a model's own performance, and low source text diversity is one primary cause.

[3] Vilém Zouhar, **Wenda Xu**, Parker Riley, Juraj Juraska, Mara Finkelstein, Markus Freitag, Daniel Deutsch. *Generating Difficult-to-Translate Texts*. On submission. MT-breaker, employs a LLM to iteratively generate difficult yet natural prompt examples to build more challenging translation benchmarks.

[4] **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.

[5] **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%).

[6] **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 study to define and quantify the bias exhibited by LLMs when assessing their own generated outputs.

[7] **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 efficient inference time optimization technique that iteratively refines the outputs of the PaLM 2 model at text span-level, achieving improvements of 1.7 MetricX on translation, 8.1 ROUGE-L on ASQA, and 2.2 ROUGE-L on topical summarization.

[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*. TACL 2024. This work provides a summary and analysis of self-correction techniques.

[9] **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 autorater, trained on synthetic data, surpasses unsupervised metrics, including those based on 175B GPT-3 and GPT-4.

[10] Chinmay Dandekar, **Wenda Xu**, Xi Xu, Siqi Ouyang, Lei Li. *Translation Canvas: An Explainable Interface to Pinpoint and Analyze Translation Systems*. EMNLP Demo 2024. System demonstration of InstructScore.

[11] **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.

[12] **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.

[13] **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 at text style transfer by 106%.

INDUSTRY RESEARCH EXPERIENCE

Full-time Research Scientist

Google Translate Research

April 2025 – Now Mountain View, CA

- **Publication:** Searching for Difficult-to-Translate Test Examples at Scale
- **Publication:** Deconstructing Self-Bias in LLM-generated Translation Benchmarks
- **Publication:** Generating Difficult-to-Translate Texts
- Developed and scaled an automated pipeline to generate dynamic challenge sets for evaluating Gemini models across both pre-training and post-training stages.
- Led the research efforts in automatic benchmark construction using LLMs, studying both self-bias in LLM-generated benchmark and efficient auto-challenge set construction.
- Led the research efforts in studying length bias of translation evaluation metric (On submission publication).

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
- Built a generic KD framework that generalizes to on-policy and supervised KD, achieves substantial gains in task specific and task agnostic knowledge distillation (now deployed in production at Google Translate).

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
- Developed an efficient, inference-time optimization technique that iteratively refines PaLM 2 outputs at the span level, which has been successfully deployed in production by the YouTube team.

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
- Developed a learned evaluation metric without human labels that achieved high correlation with human judgment and greatly improves the assessment of translation quality.

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)