Xiaopu Wang

State College, PA • xmw5221@psu.edu
shooof.github.io/homepage • github.com/shooof

Education

Pennsylvania State University 2024 – Present
Ph.D. in Statistics *University Park, PA*

University of Science and Technology of China (USTC) 2021 – 2024
M.S. in Statistics *Hefei, China*

Nankai University 2017 – 2021
B.S. in Mathematics *Tianjin, China*

Research Experience & Publications

ICML 2026 • first author 2024 – 2026
Beyond Heuristic Tuning: Power-Calibrated LLM Watermarking
Wang, X., He, Z., Liu, C., Li, R.

- Derived closed-form relationships linking logit-based watermark parameters to detection power and KL distortion, replacing heuristic tuning with a constrained optimization procedure.
- Theory matched empirical green-token rates with $R^2 > 0.98$ across all model–dataset pairs; calibrated operating points are Pareto-optimal, sustaining near-saturated detection power where heuristically tuned KGW baselines degrade sharply.
- Built the PyTorch/Hugging Face evaluation harness; benchmarked GPT-2, OPT, Pythia, and Gemma-2 9B on C4, LFQA, and Wikipedia with KL, BLEU/ROUGE, and BERTScore metrics. Code: github.com/shooof/wm.

Submitted, ICLR 2027 • first author 2026 – Present
ByteMark: A Unified Byte-Level Watermark for Detection Across Language Models
Wang, X., Li, R.

- Designed a shared UTF-8 watermark for unknown-source detection, deriving finite-sample false-positive thresholds for a public scoring rule over context–byte pairs that is verifiable from text and a key alone (no model or tokenizer access).
- Implemented the Hugging Face generation integration and a standalone detector; evaluated representation mismatch, cross-model continuation, and participating-model rewriting with watermark re-embedding.

Nature Communications 2025 • co-first author 2023 – 2025
UKB-MDRMF: A Multi-Disease Risk and Multimorbidity Framework Based on UK Biobank Data
*Jiang, Y., *Zhao, B., ***Wang, X.**, *Tang, B., *et al.*

- Built the end-to-end cleaning, feature-engineering, and survival-analysis pipeline (R + Python) over UK Biobank ($\sim$500K participants) for multi-disease risk prediction.
- Improved atrial fibrillation AUC by **0.05** and survival c-index by **0.15** (CVD) and **0.09** (CHD) over prior models.

Manuscript in preparation 2025 – Present
Marked Hawkes Processes for Recurrent Events in Electronic Health Records
Wang, X., *et al.*

- Building marked temporal point-process (deep Hawkes) models for disease risk and patient trajectories; own the cohort-construction, preprocessing, and evaluation pipelines on MIMIC-IV.

Statistics and Its Interface 2024 • first author 2021 – 2024
Copy Number Variation Detection Based on Constraint Least Squares
Wang, X., Wang, X., Zhang, A., Wen, C.

- Constrained least-squares formulation with an alternating minimization algorithm; lowest mean absolute deviation across four simulated noise settings (incl. heavy-tailed); tested on glioblastoma and HapMap data.

Teaching

Teaching Assistant 2024 – Present
Pennsylvania State University, Department of Statistics *University Park, PA*

- STAT 414/415 (Probability Theory, Mathematical Statistics); STAT 200: led and lectured lab sections.

Technical Skills

Programming: Python, R, C++, SQL
ML & LLMs: PyTorch, Hugging Face Transformers, scikit-learn, XGBoost; LLM watermarking & AI-text detection, retrieval-augmented generation, transformers, diffusion models
Statistics: hypothesis testing, survival analysis, point processes, causal inference, high-dimensional inference, constrained optimization
Data: large-scale biomedical pipelines (UK Biobank 500K, MIMIC-IV, All of Us), Linux/shell, Git