Shuxian "Trinity" Fan

2808 Calder AVE NE • Redmond, WA 98052 • fansx@uw.edu • (206) 498-6174

PROFESSIONAL EXPERIENCE

AMERICAN FAMILY INSURANCE **Seattle, WA**
Machine Learning Scientist **06/2022-09/2022**
Intern **06/2023-09/2023**

- ***Developed Model-Based System to Detect Large Language Model (LLM) Hallucinations***
  - Developed a model-based scoring and reference system to detect and quantify hallucinations in LLM outputs for automated insurance claims processing, improving claim validation accuracy by 5%.
  - Utilized in-context learning techniques to benchmark outputs against reference data, ensuring consistency and reducing misinformation rates.
- ***Enhanced out-of-distribution (OOD) Detection in Automating Medical Billing Using Transformers***
  - Enhanced OOD detection in medical billing by engineering the LayoutLM architecture, optimizing document comprehension with innovative cross-modal matching loss for automating medical bill filing and processing.
  - Improved data extraction efficiency by processing 1000+ complex medical documents weekly, reducing error rates to less than 3%.

UNIVERSITY OF WASHINGTON **Seattle, WA**
Research Assistant **05/2021-Present**

- ***Valid Inference with Prediction-Powered Inference (PPI) for LLM-Driven Verbal Autopsy (VA) Narratives***
  - Extended PPI framework for multinomial classifications, enhancing the reliability of AI-generated outcomes in downstream inference tasks, particularly in cause of death (COD) prediction from narratives.
  - Designed and implemented a robust data analysis pipeline to improve model accuracy and robustness, ensuring consistent COD modeling even in the presence of incomplete or noisy data.
  - Leveraging PPI to address AI model collapse by introducing parameter recalibration techniques, effectively mitigating the compounding of errors when models are trained on recursively generated data.
- ***Bayesian Model for Joint Analysis of Classified Data in VA Studies***
  - Designed and implemented Bayesian models to jointly analyze fully and partially classified datasets, improving data integrity by maintaining a consistent structure.
  - Applied these models to standardize age categories in VA data for under-five mortality studies, leading to up to 20% improvement in estimation accuracy, directly influencing WHO health policies and contributing to more accurate global health reporting and targeted interventions.
- ***Bayesian Active Learning for Enhanced Child Mortality Data Collection***
  - Designed and implemented Bayesian active learning strategies to refine VA questionnaire designs, shortened average questionnaire length by 20% while enhancing accuracy in child mortality assessments.

UNIVERSITY OF BRITISH COLUMBIA **Vancouver, BC**
Research Assistant **09/2018-08/2020**
Statistical Consultant **09/2019-08/2020**

- Collaborated with researchers from diverse universities and industries to tackle complex statistical challenges, delivering targeted analysis that enhanced project outcomes for over 10 high-profile studies in two years.

Selected Project Experience

- ***Enhanced Knot Detection in Timber via Modified Faster R-CNN***
  - Led original research on timber knot detection, increasing detection rate by 8%, through the use of blob detection methods in Java and processing tracheid effect data in Python.
  - Pioneered a novel approach for detecting knots in color images of sawn timber using a modified Faster R-CNN with a Gaussian Proposal Network in PyTorch to identify elliptical knot forms, and constructed 3D knot volumes for integration into timber strength modeling.
- ***Bayesian Modeling of Timber Strength Using Knot Distribution***
  - Proposed and implemented a Bayesian hierarchical model in R to characterize timber tensile strength.

- Improved predictive performance by 5% over baseline models and contributed to more accurate nationwide timber grading standards in Canada.

FPINNOVATIONS **Vancouver, BC**
Data Scientist Intern **05/2019-08/2019**

- ***Spatial Statistics for Timber Strength Analysis***
  - Collaborated with the structural engineering team to develop statistical plans for analyzing the mechanical properties of sawn timber, ensuring robust data collection and analysis strategies for improved accuracy and reliability.
  - Conducted independent research using spatial statistics to characterize timber tensile strength, designing a pilot study and analyzing large-scale industrial data sets with R and Python, resulting in enhanced predictive models and improved material strength predictions.

EDUCATION

UNIVERSITY OF WASHINGTON **Seattle, WA**
The Doctor of Philosophy in Statistics (3.8/4.0); Advanced Data Science and ML Track **2020-2025**

- Graduate Student Representative of the Department of Statistics

UNIVERSITY OF BRITISH COLUMBIA **Vancouver, BC**
Master of Science in Statistics (4.0/4.0); distinction with honors **2018-2020**

- Core organizer of the UBC/SFU Joint Statistical Seminar
- Awards: Rick WHITE Memorial Award 2020 (only 2 awarded to class)

UNIVERSITY OF BRITISH COLUMBIA **Vancouver, BC**
Bachelor of Science in Statistics (4.0/4.0); valedictorian **2015-2018**

- Awards: Stanley W. Nash Medal in Statistics 2018 (only 1 awarded to class), Dr. John and Barbara PETKAU Scholarship 2017 (first recipient), Trek Excellence Scholarship 2016, 2017 (top 5%)

SELECTED PUBLICATIONS

- **Fan, Shuxian**, et al. "From Narratives to Numbers: Valid Inference Using Language Model Predictions from Verbal Autopsies." First Conference on Language Modeling.
- Yoshida, Toshiya, **Fan, Shuxian**, et al. "Bayesian Active Questionnaire Design for Cause-of-Death Assignment Using Verbal Autopsies." Conference on Health, Inference, and Learning. PMLR, 2023.
- **Fan, Shuxian**, Samuel WK Wong, and James V. Zidek. "Knots and their effect on the tensile strength of lumber: A case study." Journal of Quality Technology 55.4 (2023): 510-522.
- **Fan, Shuxian**, et al. "Ellipse detection and localization with applications to knots in sawn lumber images." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2021.

RELEVANT SKILLS

- ***Technical Skills***
  - Programming: Python, SQL, R, SAS, Java, C++, Julia, MATLAB
  - Data Visualization: Matplotlib, NumPy, Pandas, ggplot2, Seaborn, D3.js, Tableau,Shiny app
  - ML and AI: PyTorch, TensorFlow, Hugging Face Transformers, OpenAI API, spaCy, AllenNLP, scikit-learn, PyCaret, Keras, XGBoost
  - Computing: Spark, Databricks, Hive, AWS, Azure, Google Cloud, Git, GitHub, Bitbucket
- ***Research and Statistical Skills***
  - Regression, Clustering, Classification, Hypothesis testing, A/B testing, Time Series Analysis, Stochastic Processes, Experimental Design, Bayesian Nonparametrics, Active and Reinforcement Learning, Optimization
  - Text Processing (Tokenization, stemming, lemmatization, POS tagging, and NER), Text Generation, Classification and Summarization (GPT, BERT, RoBERTa), Document Understanding (LayoutLM) Computer Graphics and Sequence Modeling (CNNs, RNNs, GANs, Transformers, Autoencoders)