Yennie Jun

31-84 30th St, Astoria, NY | yennie.jun@gmail.com | 225·290·6216 | blog.yenniejun.com

WORK EXPERIENCE

Truveta Remote
Senior Machine Learning Engineer *Feb. 2022—Present*

- Train open-source Transformer models on millions of electronic health records for clinical concept detection, Personal Health Information (PHI) redaction, and relation extraction, resulting in state-of-the-art F1 score on public datasets
- Implement active learning for extraction models, reducing human annotation cost by 3x while achieving over 95% recall
- Pre-annotate documents using weak supervision to improve human annotators' productivity and efficiency
- Lead efforts to leverage closed-source generative language models for generating internal coding language and clinical concept codes to aid clinical researchers in conducting studies

DeepLearning.AI Remote
Consultant, Technical Writer *Aug. 2021—Present*

- Write articles for The Batch on state-of-the-art machine learning research such as reinforcement learning, language models, audio generation models, and NeRF
- Designed curriculum and created teaching materials for online course on AI for social good

United Nations (Global Pulse) New York, NY, USA
Data Scientist *Apr. 2021—Feb. 2022*

- Led speech-to-text benchmarking project comparing several commercial and open-source ASR models, including organizing annotation process of 600+ radio segments from 5 African countries
- Conducted NLP analyses on 4000+ hours of public radio transcriptions about COVID-19 public health concerns
- Built custom dashboard for querying, indexing, and surfacing radio transcriptions for WHO public health officials

Seoul National University
Research Data Scientist Seoul, South Korea

- Gathered and cleaned COVID-19 news data from 6 Asian countries in 3 languages to analyze contact tracing discourse early in the pandemic using topic modeling and semantic network analysis *Nov. 2019—Apr. 2021*
- Performed entity recognition, resolution, and disambiguation of 14K historical Korean civil service figures

Microsoft Redmond, WA, USA
Software Engineer *Sep. 2017—Sep. 2019*

- Full-stack engineer on the Microsoft Education platform. Integrated existing learning management system with Immersive Reader, an AI-powered tool for enhancing reading comprehension

PROJECTS

- **OpenAI Red Team.** Part of the red team of AI researchers probing DALLE-2 and GPT-4 for potentially harmful and dangerous generated images prior to its public release. Advised and created recommendations for mitigating harmful effects and risks of the model.
- **Art Fish Intelligence**. Created a personal blog dedicated to exploring the behavior of language models and generative AI through data analysis and experimentation. Topics included prompt engineering, code generation, AI-assisted creative writing, and analyzing my personal health data
- **COVID-19 Vaccine Discourse on Public Radio in South Africa and Nigeria**. Used temporal word embeddings to evaluate COVID-19 vaccine discourse on public radio streams from South Africa and Nigeria. Analyzed word embedding stability through running hundreds of models. Master's Thesis with University of Oxford.
- **Bias Out-of-the-Box: An Empirical Analysis of Intersectional Occupational Biases in Popular Generative Language Models**. Probed GPT-2 with prefix templates related to gender and occupation to evaluate biases in its predictions, which were compared with ground-truth US labor data. NeurIPS, 2021 (arXiv:2102.04130).
- **Chatbot for COVID-19 Info**. Created and deployed texting service for answering questions related to the pandemic and providing COVID-19 statistics to those without access to Internet. Reached 100+ daily users. Acquired by Silicon Harlem.

EDUCATION

University of Oxford Oxford, United Kingdom
MSc in Social Data Science (Distinction), Oxford Internet Institute *Oct. 2020—Aug. 2021*

Tufts University Medford, MA, USA
BS in Computer Science, BA in History *Sep. 2013—May 2017*

TECHNICAL SKILLS

Languages: Python, JavaScript, C/C++/C#, SQL, HTML/CSS
Machine Learning: HuggingFace Transformers, NLTK, NumPy, Pandas, PyTorch, scikit-learn, spaCy, TensorFlow, gensim
Software: Jupyter, conda, Dash/Plotly, Flask, Git, Postman, React.js, Twilio