Jenny Chim

✉ jennychim@gmail.com 🔗 jennycwchim 📘 Kuhep28AAAAJ

EDUCATION

Ph.D. Computer Science, Queen Mary University of London   2020 – present
- Researching synthetic data, language generation and large language model evaluation.
- Organising the Data Science for Mental Health interest group at the Alan Turing Institute.

M.Sc. Speech and Language Processing (distinction), The University of Edinburgh   2018
- Thesis: "Exploring neural networks for named entity recognition in radiology".

B.A. Psychology (with honors), University of California, Los Angeles   2017
- Minor: Linguistics

EXPERIENCE

Research Engineer (Contract), Anathem   2023 – present
- Designing and implementing model evaluation and MLOps framework.

Data Scientist, EF Education First   2018 – 2020
- Developed machine learning models and reports on the automatic classification of course content, user behaviour, teacher feedback, and student responses.
  - Utilised: *scikit-learn, PyTorch, Tensorflow; Flask; Docker; streamlit.*
- Supported implementation and maintenance of data ingestion pipeline on AWS.
  - Utilised: *Airflow; Snowflake; Kafka, Spark.*
- Built interactive dashboards to support reporting and real-time analytics.
  - Utilised: *NumPy, pandas, matplotlib, seaborn, D3.JS.*

Human Testing Technician, University of Edinburgh   2018
- Managed data collection, stimuli processing, and subject recruitment for eye-tracking experiments studying infant metacognitive sensitivity.

Research Assistant, The Hong Kong Polytechnic University   2017
- Responsible for subject recruitment, IPA transcription, experiment scripting, and acoustic analyses.

Undergraduate Research Assistant, University of California, Los Angeles   2016 – 2017
- *Language Processing Lab*: Annotated corpora, prepared audio stimuli, and conducted psycholinguistics experiments (eye-tracking, self-paced reading, lexical decision, questionnaires).
- *Language and Cognitive Development Lab*: Recruited participants, collected data, and coding.

PUBLICATIONS

NLP for Mental Health
- **Chim, J.***, Song, J*., et al. (2024). Combining Hierachical VAEs with LLMs for clinically meaningful timeline summarisation in social media. *arXiv:2401.16240.*
- **Chim, J***, Tsakalidis, A*., et al. (2024). Overview of the CLPsych 2024 Shared Task: Leveraging Large Language Models to Identify Evidence of Suicidality Risk in Online Posts.
- Tsakalidis, A., **Chim, J.**, et al. (2022). Overview of the CLPsych 2022 Shared Task: Capturing Moments of Change in Longitudinal User Posts.
- Tsakalidis, A., [et al, including **Chim, J**] (ACL 2022). Identifying Moments of Change from Longitudinal User Text.

Large Language Models
- Scao, T. L., [et al, including **Chim, J**] (2022). Bloom: A 176b-parameter open-access multilingual language model. *arXiv:2211.05100.*
- Laurençon, H., [et al, including **Chim, J**] (NeurIPS Datasets and Benchmarks 2022). The bigscience roots corpus: A 1.6 tb composite multilingual dataset.
- Li, R., [et al, including **Chim, J**] (TMLR 2023). StarCoder: may the source be with you!.

Data & Evaluation Frameworks
- Gehrmann, S., [et al, including **Chim, J**] (EMNLP 2022). GEMv2: Multilingual NLG Benchmarking in a Single Line of Code.
- Fries, J., [et al, including **Chim, J**] (NeurIPS Datasets and Benchmarks 2022). Bigbio: a framework for data-centric biomedical natural language processing.

SKILLS

Programming/Scripting: Python, Scala, SQL, R, Praat
Languages: *Native*: English, Cantonese, Mandarin Chinese; *Working proficiency*: Japanese