JENNY CHIM

jennychim@gmail.com | linkedin.com/in/jennycwchim | Google Scholar

SUMMARY

Applied scientist with 5+ years working on LLM evaluation, synthetic data, and applied NLP across healthcare, education, and data engineering. Experienced in building and assessing production-facing models, designing evaluation frameworks, and developing datasets in partnership with domain specialists. Strong cross-functional collaborator with a track record of translating research into reliable tooling and workflows.

TECHNICAL SKILLS

Languages: Python (primary), SQL, Scala, R, Bash
ML/NLP: PyTorch, TensorFlow, Scikit-learn, Transformers, Pandas, NumPy, spaCy, nltk
LLM Tooling: LangChain, LlamaIndex, LangFuse
Platforms & Tools: Docker, AWS (SageMaker, EC2, S3), Azure, Snowflake, Airflow, Streamlit
Evaluation: dataset construction, human-in-the-loop workflows, rubric development, privacy/memorization audits, hallucination/reliability analysis, prompt optimization, multi-turn agent evaluation

EXPERIENCE

AI Scientist, *Matillion* — *2025 – Present*

- Designing and implementing internal agent evaluation benchmarks for conversational quality and long-horizon tasks.
- Developing error-clustering and log-mining methods to improve measurability and strengthen evaluation coverage.

Postdoctoral Researcher, *Queen Mary University of London* — *2024 – 2025*

- Researched model evaluation in the RAi UK Keystone project AdSoLve, focusing on medical and legal applications.
- Developed question answering and summarization evaluation methods (NLI-based scoring, LLM-as-a-judge, fact decomposition), focusing on grounding, consistency, and alignment with domain experts.

Lead Interest Group Organiser, *The Alan Turing Institute* — *2021 – 2025*

- Led day-to-day operations for the Data Science for Mental Health (DS4MH) special interest group.
- Managed event planning, membership management, and outreach strategies to grow the research community.

Data Scientist, *NHS England (PhD Internship)* — *2024 – 2024*

- Researched privacy concerns and mitigations, and contributed to internal reporting for LLM monitoring.
- Developed privacy risk evaluation suites, focusing on memorization in fine-tuned models (prompt-based extraction, automatic red teaming) and information flow in LLMs for ambient scribing (synthetic data, semantic search).

Research Engineer, *Anathem (Consultant)* — *2023 – 2024*

- Led research and implementation on evaluation for AI-supported clinical documentation in mental health.
- Implemented document generation features using LangChain and user analytics dashboard on Streamlit.

Data Scientist, *EF Education First* — *2018 – 2020*

- Developed machine learning models for content analysis, user modeling, and business analytics.
- Implemented and maintained data ingestion and governance pipelines on AWS using Airflow and Snowflake.
- Deployed interactive internal tools using Docker and Streamlit to visualize model outputs for stakeholders.

EDUCATION

Queen Mary University of London *Ph.D. in Computer Science* — *2020 – 2025*

- **Thesis:** Synthetic Data and Evaluation Methods for Longitudinal Language Processing.
- **Funding:** DeepMind PhD Scholarship (2020 – 2024).

The University of Edinburgh *M.Sc. Speech and Language Processing (Distinction)* — *2018*

- **Thesis:** Deep Learning Methods for Named Entity Recognition in Radiology.

University of California, Los Angeles (UCLA) *B.A. Psychology, Minor in Linguistics (Phi Beta Kappa)* — *2017*

- **Research Experience:** Supported projects at the Language Processing Lab, Language and Cognitive Development Lab, and Political Science department by managing data collection, experiment scripting, and data analyses.

Selected Papers

Full publication list available on Google Scholar.

Large Language Models

- Li, R., [et al, incl **Chim, J**] (TMLR 2023). StarCoder: may the source be with you!
- Scao, T. L., [et al, incl **Chim, J**] (2022). Bloom: A 176b-parameter open-access multilingual LM.
- Laurençon, H., [et al, incl **Chim, J**] (NeurIPS D&B 2022). The BigScience Roots Corpus.

Evaluation Frameworks

- **Chim, J.**, Ive, J., Liakata, M. (Computational Linguistics 2025). Evaluating Synthetic Data Generation from User Generated Text.
- Vayani, A., [et al, incl **Chim, J**] (CVPR 2025). All Languages Matter: Evaluating LMMs on Culturally Diverse 100 Languages.
- Zhuo, T.Y., Vu M.C., **Chim, J.**, et al. (ICLR 2025). BigCodeBench: Benchmarking Code Generation with Diverse Function Calls and Complex Instructions.
- Romanou, A., [et al, incl **Chim, J**] (ICLR 2025). INCLUDE: Evaluating Multilingual Language Understanding with Regional Knowledge.
- Fries, J., [et al, incl **Chim, J**] (NeurIPS D&B 2022). BigBio: a framework for data-centric biomedical natural language processing.
- Gehrmann, S., [et al, incl **Chim, J**] (EMNLP 2022). GEMv2: Multilingual NLG Benchmarking in a Single Line of Code.

Responsible AI and Applications

- **Chim, J.,*** Ghosh, A*., Reuel A*., et al. (Under Review). Who Evaluates AI's Social Impacts? Mapping Coverage and Gaps in First and Third Party Evaluations.

NLP for Healthcare

- **Chim, J.,** et al. (Under Review). Evaluating Privacy Leakages in LLM-driven Ambient Clinical Documentation.
- AlMannaa F, [et al, incl **Chim, J**] (Under Review). Investigating LLM Capabilities on Long Context Comprehension for Medical Question Answering.
- **Chim, J.***, Tseriotou, T*., et al. (2025). Overview of the CLPsych 2025 Shared Task: Capturing Mental Health Dynamics from Social Media Timelines.
- **Chim, J.***, Song, J*., et al. (ACL Findings 2024). Combining Hierarchical VAEs with LLMs for clinically meaningful timeline summarisation in social media.
- **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, incl **Chim, J**] (ACL 2022). Identifying Moments of Change from Longitudinal User Text.

Selected Presentations

Talks

- Spring 2025. Evaluating Privacy Leakages in LLM-driven Ambient Clinical Documentation. HealTAC.
- Fall 2024. Evaluating Synthetic Data Generation from User Generated Text. EMNLP (Oral).
- Fall 2024. Privacy Concerns and Mitigations for Healthcare Language and Foundation Models. Privacy Enhancing Technologies Cross-Government (xGov) Meeting.
- Spring 2023. BigCode: StarCoder Model Review. Webinar.

Tutorials

- Spring 2023. Prompt Engineering. NLP Interest Group, The Alan Turing Institute.

Contributed Presentations

- Summer 2024. Combining Hierachical VAEs with LLMs for clinically meaningful timeline summarisation in social media. ACL (Poster).
- Spring 2022. Identifying Moments of Change from Longitudinal User Text. ACL (Poster).

Languages

Native: English, Cantonese, Mandarin Chinese | **Working Proficiency**: Japanese