Chenghui Zhou

https://zhouchenghui.github.io/
+44 20 3996 2697
chenghui.zhou25@gmail.com

Profile

Applied Scientist at Microsoft specializing in generative models, multimodal models, and NLP. Experienced in training and evaluating LLMs and VLMs. Eligible to work in the UK without sponsorship.

Education

Ph.D. in Machine Learning – Carnegie Mellon University — July 2024
Thesis Topic: Generative Models for Structured Discrete Data with Application to Drug Discovery
Advisor: Barnabás Póczos

M.S. in Machine Learning – Carnegie Mellon University — December 2018
Related Coursework: Deep Reinforcement Learning, Probabilistic Graphical Models, Intermediate Statistics, Statistical Machine Learning, Convex Optimization

B.S. in Honours Computer Science – McGill University — May 2016
Minor in Statistics
Advisor: Joelle Pineau

Skills

Programming Languages: Python, Java, C, Matlab, LaTeX
Platforms & Technologies: Azure, AWS, Git, Linux, Windows, OS X
Machine Learning Packages: PyTorch, JAX, TensorFlow, scikit-learn, Hugging Face Trasnformers, DeepSpeed
Languages: English (*Fluent*), Chinese (*Fluent*), German (*Beginner*), French (*Beginner*)

Industry & Research Experience

Applied Scientist II — July 2024 – Present
Microsoft AI Development Acceleration Program — Boston, USA
In collaboration with Microsoft MSAI Turing Team (December 2025 – present)
- Applying and evaluating post-training techniques to masked diffusion models and LLM multi-token generation for coding and math tasks to improve inference efficiency.

In collaboration with Microsoft Office AI (April 2025 – December 2025)
- Fine-tuned and benchmarked a multimodal system (Phi-3.6/SigLIP2) for PowerPoint Copilot visual summarization and rewriting features optimized for on-device inference.
- Reduced the performance gap between the multimodal system and GPT-4.1 by 62% across iterations (0.8→0.3 on a 5-pt LLM evaluation scale) on the target PowerPoint dataset.

In collaboration with Azure Health Service (January 2025 – June 2025)
- Analyzed correlations with outages on an Azure service health database with millions of time series
- Developed a foundation-model-based outage detection framework for service health monitoring

In collaboration with Microsoft Teams (July 2024 – December 2024)
- Fine-tuned BERT for real-time phishing detection in Chat using LLM-generated synthetic data.
- Improved precision over the best baseline by 30%, with the model deployed to production.

Applied Scientist Intern — May 2020 – August 2020
CodeGuru team, Amazon Web Service, Inc. — Pittsburgh, USA
- Developed a contrastive learning method by leveraging CodeBERT embeddings for effective code retrieval from repositories

Research Intern — May 2018 – August 2018
Predictive Algorithm team, Zoll LifeVest — Pittsburgh, USA
- Designed and trained sequence models for robust classification of cardiac rhythms in ECG segments

Chenghui Zhou, chenghui.zhou25@gmail.com

Undergraduate Research Assistant May 2014 – August 2015
Reasoning and Learning Lab, McGill University Montreal, Canada
- Developed a predictive linear Gaussian algorithm to improve predictions for tracking in robotics
- Evaluated the algorithm on synthetic and real data of human walking trajectories

Publications

- Yuchen Shen*, Chenhao Zhang*, Sijie Fu*, **Chenghui Zhou**, Newell Washburn, Barnabás Póczos
  Chemistry-Inspired Diffusion with Non-Differentiable Guidance
  International Conference on Learning Representations (ICLR), 2025

- **Chenghui Zhou***, Yuchen Shen*, Chenhao Zhang*, Sijie Fu, Newell Washburn, Barnabás Póczos
  Non-Differentiable Diffusion Guidance for Improved Molecular Geometry
  AI4Science Workshop, ICML, 2024.

- **Chenghui Zhou**, Barnabás Póczos
  Improving Molecule Properties Through 2-Stage VAE
  Machine Learning for Structural Biology Workshop, NeurIPS, 2022.

- **Chenghui Zhou***, Frederic Koehler*, Viraj Mehta*, Andrej Risteski
  Variational Autoencoders in the Presence of Low-Dimensional Data: Landscape and Implicit Bias
  International Conference on Learning Representations (ICLR), 2022.

- **Chenghui Zhou**, Chun-Liang Li, Barnabás Póczos
  Unsupervised Program Synthesis for Images by Sampling without Replacement
  Conference on Uncertainty in Artificial Intelligence (UAI), 2021.

- **Chenghui Zhou**, Chun-Liang Li, Barnabás Póczos
  Unsupervised Program Synthesis for Images Using Tree-Structured LSTM
  Deep Reinforcement Learning Workshop & Learning with Rich Experience Workshop (one of two selected oral presentations) *NeurIPS*, 2019.

- Robin Schmucker, **Chenghui Zhou**, Manuela Veloso
  Multimodal Movement Activity Recognition Using a Robot's Proprioceptive Sensors
  RoboCup Symposium, 2018.

- **Chenghui Zhou**, Manuela Veloso
  Interception in Continuous Space Using Deep Reinforcement Learning
  Submitted to *International Conference on Robotics and Automation (ICRA)*, 2018.

- Michiel de Jong, Kevin Zhang, Travers Rhodes, Aaron Roth, Robin Schmucker, **Chenghui Zhou**, Sofia Ferreira, João Cartucho, Manuela Veloso
  Towards a Robust Interactive and Learning Social Robot
  International Conference on Autonomous Agents and Multiagent Systems (AAMAS), 2018.

- **Chenghui Zhou**, Borja Balle, Joelle Pineau
  Learning Time Series Models for Pedestrian Motion Prediction
  International Conference on Robotics and Automation (ICRA), 2016.

Teaching

2021 Teaching assistant for **Convex Optimization** (Carnegie Mellon University)

2018 Teaching assistant for **Statistical Machine Learning** (Carnegie Mellon University)

2016 Teaching assistant for **Introduction to Software Systems** (McGill University)

2015 Teaching assistant for **Foundations of Programming** (McGill University)