

XUETONG WU

Staff AI Engineer at Heidi Health | Ph.D. in Electrical Engineering | LLM & ASR | Information Theory

📍 Melbourne, Australia ✉️ wfyitf@gmail.com 🌐 wuxuetong.com 🔗 linkedin.com/in/xuetong-wu 🎓 [Google Scholar](https://scholar.google.com/citations?user=...)

PROFILE

Staff AI Engineer at Heidi Health, building the fine-tuned large language models and automatic speech recognition systems behind Heidi's clinical scribe, which replaced closed frontier models in production, and tech-leading the on-premises deployment of its clinical AI stack. Ph.D. in Electrical Engineering from the University of Melbourne with 10+ first- or second-author papers in IEEE Transactions on Information Theory, JMLR, Artificial Intelligence, ACL, and NAACL, and recipient of the 2023 Best EEE Ph.D. Thesis Prize. Work spans information-theoretic foundations, fine-tuning, evaluation design, and latency- and cost-constrained serving.

PROFESSIONAL EXPERIENCE

Heidi Health | Staff AI Engineer

2026.08 – Present

- **On-premises Deployment:** Tech lead for an air-gapped deployment of Heidi's clinical AI stack (transcription, note generation, dictation, translation, evidence retrieval) for hospitals with data-sovereignty requirements, served entirely on local NVIDIA GPUs across 8 languages.
- **Multilingual ASR:** Distil commercial teachers into self-hosted Whisper and Qwen students; the Korean model reached parity with the commercial cloud engine in blind clinician review, and a single Arabic/Urdu/English model outperformed its teacher on clinical Arabic.
- **Evidence & Forms:** Built an offline clinical evidence engine with cited answers (BM25, embedding scoring, cross-encoder reranking), and structured form filling with schema-constrained decoding that eliminated unparseable outputs.
- **Leadership:** Mentor engineers across the Models team, interview senior and staff candidates, and lead technical discovery with enterprise customers.

Heidi Health | Senior AI Engineer

2025.10 – 2026.08

- **Note Generation:** Delivered fine-tuned LLMs (SFT, DPO, distillation) that replaced expensive closed models in production, saving millions of dollars per year.
- **Transcription:** Led the workstream; fine-tuned an in-house clinical ASR model (Parakeet) that took over most English production traffic from vendor APIs at lower latency and cost.
- **Dictation:** Showed 93% of quality survives 3-second chunks (down from 30 s), moving production to near-real-time dictation with a sub-second P95 target.
- **Infrastructure:** Stood up the first on-premises GPU servers running Heidi's scribe models, with Triton and vLLM serving load-tested on NVIDIA GB10.
- **Agentic Evidence:** Built the trajectory-distillation pipeline for a clinical evidence model that searches and reasons across multiple tool calls.

Heidi Health | AI Engineer

2025.06 – 2025.10

- **Post-training Data:** Built the note-generation data pipeline: sampled production sessions across five regions, regenerated targets and filtered them with LLM judges before SFT and DPO.
- **Model Comparison:** Ran side-by-side experiments between fine-tuned variants and frontier models on live traffic, tracking clinician satisfaction and edit rates.
- **Shipping:** Released an improved clinical code generation model; contributed to a collaboration with NVIDIA on fine-tuning speech recognition for clinical use.

StrongRoom AI | Machine Learning Engineer

2023.06 – 2025.05

- **Business Impact:** Drove platform optimisations improving medication adherence tracking across 400+ pharmacies, contributing to AUD 65M in revenue with AUD 1.5M attributed directly to improved prediction models.
- **NLP Engineering:** Built pipelines extracting structured data from unstructured medical records, reaching 95%+ accuracy on dosage extraction.
- **Computer Vision:** Developed embedded systems for real-time facial analysis in opioid dependence studies with Alfred Hospital, achieving 90%+ accuracy in detecting early signs of medication non-adherence.
- **Infrastructure:** Designed scalable ETL systems processing multi-terabyte healthcare datasets across distributed pharmacy networks.
- **Collaboration:** Worked with data scientists, clinicians, and engineers to align model behaviour with clinical and operational requirements.

- Developed the StrongPro Platform for pharmacy owners and medical institutions, targeting patient adherence and loyalty prediction.
- Analysed and visualised medical data to surface insights on patient behaviour, pharmacy revenue, and recommended interventions.
- Built the initial NLP and ML pipelines for medical data extraction and adherence computation.

The University of Melbourne | Research Fellow, Applied AI

2024.10 – 2025.02

Advanced Evaluations for L2 English Applications with Large Language Models

- Researched LLM-based evaluation methodology for second-language (L2) English learners, covering both quantitative metrics and qualitative feedback.
- Built an automation framework for interactive evaluation of Chinese-as-Second-Language dialogues, published at NAACL HLT 2025 (oral, top 15%).
- Quantified how faithfully LLMs simulate L2 English learners across L1 backgrounds, published at ACL 2025 Main Conference.

The University of Melbourne | Research Fellow, Multi-modal AI

2024.01 – 2024.10

Real-time Multi-modal Out-of-Distribution Detection in Extended Dialogue Scenarios

- Developed frameworks for real-time multi-modal out-of-distribution detection in long-form dialogue, with novel metrics reaching state-of-the-art performance.
- Combined textual and visual signals to improve robustness across conversational datasets; results published at ACL 2025 Findings.
- Coordinated a cross-disciplinary team of NLP and computer vision researchers alongside industry partners.

The University of Melbourne | Teaching Fellow, Dept. of EEE

2019.04 – 2024.12

- Ran laboratory and tutorial sessions for graduate courses of 50+ students per semester, and mentored master’s students on transfer learning projects.
- Taught ELEN90058 Signal Processing (2021–2022) and authored tutorial material for ELEN90052 Advanced Signal Processing (2019–2022).
- Highest teaching satisfaction rating in the department for ELEN90052 in 2019.

FoRECAst Study, The University of Melbourne | Data Analyst Researcher

2019.03 – 2019.12

- Built models predicting amenorrhea status following cancer treatment, with imputation and transfer learning methods that materially reduced prediction error on records containing substantial missing data.
- Published two papers on imputation and transfer learning for healthcare data.

The University of Melbourne, COMMNETS Lab | Research Intern

2018.12 – 2019.03

- Designed deep learning models for communication coding schema design, improving error correction for binary and complex codes, validated through large-scale system simulation.

TECHNICAL SKILLS

Post-training	SFT, LoRA, QLoRA, DPO, GRPO, reward modelling, RLHF, distillation
Speech	CTC, TDT, Whisper, Parakeet, Qwen-ASR, Wav2Vec2, streaming ASR, diarisation
Evaluation	LLM-as-judge, benchmark design, human-preference alignment, regression testing
Serving	vLLM, SGLang, NVIDIA Dynamo, TensorRT, ONNX, Triton, air-gapped on-premises deployment
Programming	Python (PyTorch, HuggingFace, Pandas), Rust, MATLAB, JavaScript/Vue, LaTeX
Infrastructure	Docker, Kubernetes, AWS, Airflow, Spark, FastAPI, Git
Data	MongoDB, PostgreSQL, MySQL, Delta Lake
Languages	Mandarin Chinese (native), English (professional)

EDUCATION

The University of Melbourne | Ph.D., Electrical and Electronic Engineering

2019.03 – 2023.08

- **Thesis:** Information-theoretic Analysis for Machine Learning and Transfer Learning: Bounds and Applications.
- **Supervisors:** Dr. Jingge Zhu, Prof. Jonathan H. Manton, Prof. Uwe Aickelin.
- **Coursework:** Statistical Machine Learning (H1), Optimisation (H2A), Causal Inference (H1).

The University of Melbourne | M.E., Electrical and Electronic Engineering

2016.09 – 2018.12

- **Thesis:** 3D Low-cost LiDAR Mapping for Scene Reconstruction.
- **Result:** GPA 87/100; Dean’s Honours List 2017 and 2018 (top 3%).

- **Thesis:** Pseudo-HDR Algorithm for Single Image Improvement Based on Segmentation.
- **Result:** GPA 3.22/4 (top 30%); Outstanding Student Scholarship, 2015 and 2016.

PUBLICATIONS

26 works, 100+ citations, h-index 7 (Google Scholar). * denotes co-first authorship.

- [1] **Wu, X.**, Aickelin, U., Manton, J.H. & Zhu, J. (2025). Fast Rate Information-theoretic Bounds on Generalization Errors. *IEEE Transactions on Information Theory*. (CORE A*) [arXiv:2303.14658](#)
- [2] Gao, R., **Wu, X.***, Kuribayashi, T., Ye, M., Qi, S., Roever, C., Liu, Y., Yuan, Z. & Lau, J.H. (2025). Can LLMs Simulate L2-English Dialogue? An Information-Theoretic Analysis of L1-Dependent Biases. *ACL 2025, Main Conference*. (CORE A*) [arXiv:2502.14507](#)
- [3] Gao, R., **Wu, X.***, Han, C., Luo, S. & Liu, F. (2025). “No” Matters: Dialogue-level OOD Detection with Image in Question-Answering Systems. *ACL 2025, Findings*. (CORE A*) [arXiv:2410.23883](#)
- [4] Gao, R., Wu, J., **Wu, X.**, Roever, C., Lv, L. & Lau, J.H. (2025). An Interpretable and Crosslingual Method for Evaluating Second-Language Dialogues. *NAACL HLT 2025*. (CORE A, Oral, top 15%) [arXiv:2408.16518](#)
- [5] **Wu, X.**, Gong, M., Manton, J.H., Aickelin, U. & Zhu, J. (2024). On Causality in Domain Adaptation and Semi-Supervised Learning: An Information-Theoretic Analysis for Parametric Models. *Journal of Machine Learning Research*. (CORE A*) [arXiv:2205.04641](#)
- [6] **Wu, X.**, Manton, J.H., Aickelin, U. & Zhu, J. (2024). On the Generalization for Transfer Learning: An Information-Theoretic Analysis. *IEEE Transactions on Information Theory*. (CORE A*) [arXiv:2207.05377](#)
- [7] **Wu, X.**, Manton, J.H., Aickelin, U. & Zhu, J. (2023). A Bayesian Approach to (Online) Transfer Learning: Theory and Algorithms. *Artificial Intelligence (AIJ)*. (CORE A*) [doi:10.1016/j.artint.2023.103991](#)
- [8] **Wu, X.** (2023). Information-theoretic Analysis for Machine Learning and Transfer Learning: Bounds and Applications. *Ph.D. Thesis, The University of Melbourne*. (Best EEE Ph.D. Thesis 2023) [minerva-access.unimelb.edu.au](#)
- [9] Jia, J., **Wu, X.**, Evans, J. & Zhu, J. (2023). On the Value of Stochastic Side Information in Online Learning. *IEEE ICASSP 2023*. (CORE B) [arXiv:2303.05914](#)
- [10] **Wu, X.**, Manton, J.H., Aickelin, U. & Zhu, J. (2022). Fast Rate Generalization Error Bounds: Variations on a Theme. *IEEE Information Theory Workshop 2022*. (Flagship) [IEEE Xplore](#)
- [11] **Wu, X.**, Manton, J.H., Aickelin, U. & Zhu, J. (2021). Online Transfer Learning: Negative Transfer and Effect of Prior Knowledge. *IEEE International Symposium on Information Theory 2021*. (Flagship) [IEEE Xplore](#)
- [12] **Wu, X.**, Manton, J.H., Aickelin, U. & Zhu, J. (2020). Information-theoretic Analysis for Transfer Learning. *IEEE International Symposium on Information Theory 2020*. (Flagship) [IEEE Xplore](#)
- [13] **Wu, X.**, Khorshidi, H.A., Aickelin, U., Edib, Z. & Peate, M. (2020). Transfer Learning to Enhance Amenorrhea Status Prediction in Cancer and Fertility Data with Missing Values. *AI Applications in Healthcare Delivery* (book chapter). [doi:10.4324/9780429317415-13](#)
- [14] **Wu, X.**, Khorshidi, H.A., Aickelin, U., Edib, Z. & Peate, M. (2019). Imputation Techniques on Missing Values in Breast Cancer Treatment and Fertility Data. *Health Information Science and Systems*, 7(1). [doi:10.1007/s13755-019-0082-4](#)

AWARDS & SERVICE

Best EEE Ph.D. Thesis Prize 2023 , The University of Melbourne	2025
DSTG AI for Decision Making Initiative , research funding (competitive, AUD 30,000+)	2022
Dean's Honours List 2017 & 2018 , The University of Melbourne (top 3% of engineering graduates)	2018, 2019
Outstanding Student Scholarship , Fudan University (top 30% of cohort)	2015, 2016
Reviewing Outstanding Reviewer, IEEE Transactions on Information Theory; also ISIT, ICASSP, TMLR, ICML	
Talks guest lecture on LLM post-training, UniMelb ELEN90088 (2026); NAACL 2025 oral; ITW 2022; ISIT 2020 and 2021; ICASSP 2023 poster; ICHIS 2019	
Memberships IEEE, ACM, Australian EEE/Computer Society	