

Yu-Neng (Allen) Chuang

📞 (979) 334-0045 ✧ ✉ yc146@rice.edu ✧ 🌐 Homepage: <https://ynchuang.github.io>

RESEARCH INTERESTS & SKILLS

- **Large Language Models (LLMs):** Multi-modal Agentic LLMs, Multi-agent LLMs, LLM routing, LLM Reasoning for Tool Usage, LLM Efficiency (Long Context & Efficient Reasoning)
- **Trustworthy AI/LLMs:** LLM Reasoning/Explanations, LLMs Uncertainty, LLMs Safety, Explainable AI
- **Skills:** Python, C++, Java, TensorFlow, PyTorch, Model Contextual Protocol

EDUCATION

Rice University (Co-advised by Dr. Xia “Ben” Hu & Dr. Vladimir Braverman) <i>Ph.D. in Computer Science</i>	Houston, TX <i>Aug. 2021 - Present</i>
National ChengChi University <i>Master of Science in Computer Science</i>	Taipei, Taiwan <i>Feb. 2018 - Jun. 2020</i>
National ChengChi University <i>Bachelor of Science in Mathematical Science</i>	Taipei, Taiwan <i>Aug. 2013 - Jul. 2017</i>

EXPERIENCE

Google DeepMind @ Gemini Team <i>Research Intern</i>	New York City, NY <i>May 2025 - Aug. 2025</i>
<ul style="list-style-type: none">• Developed multi-modal agentic LLMs leveraging the Model Context Protocol (MCP) with reasoning for dynamic tool planning and usage in complex audio tasks, reducing system failure rates by 28% and achieving hallucination-free tool calling.	

Apple Inc. @ Siri Team <i>Research Intern</i>	Cupertino, CA <i>May 2024 - Aug. 2024</i>
<ul style="list-style-type: none">• Developed uncertainty quantification algorithm for on-device LLMs based on confidence token prediction for instance routing and rejection learning purposes, deducing 50% of instance inference latency	

Samsung Research America @ LLM Team <i>Research Intern</i>	Mountain View, CA <i>May 2023 - Aug. 2023</i>
<ul style="list-style-type: none">• Developed an efficient algorithm of hard prompt compression on large language models with LLM post-training techniques, Proximal Policy Optimization (PPO), deducing 80% of LLM API usage cost and 20% of latency of white box LLMs	

Rice University <i>Graduate Research Assistant</i>	Houston, TX <i>Aug. 2021 - Present</i>
<ul style="list-style-type: none">• Developed fine-tuning and alignment algorithms to address uncertainty, enhance explainability, and improve safety in LLMs• Investigated multi-agent systems to enhance LLM reasoning, planning, and inference routing• Built efficient framework for KV-cache lossy compression for lower inference throughput and budgets	

Carnegie Mellon University and SMU @ Living Analytics Research Centre <i>Research Assistant</i>	Singapore <i>Jan. 2020 - Apr. 2020</i>
<ul style="list-style-type: none">• Built a ranking method for a personalized job recommendation system with TB-scaled user data in Singapore, which outperformed other state-of-the-art ranking methods by 10.2%	

KKBOX Co, Ltd. <i>Data Scientist Intern</i>	Taipei, Taiwan <i>Sep. 2019 - Jun. 2020</i>
<ul style="list-style-type: none">• Developed a ranking algorithm on a streaming dataset of nearly 1.5 million users to enhance the recommendation system by 15.3% compared with the prior internal recommendation systems	

SELECTED PUBLICATIONS

Conference and Journal Publications

- [ICML' 25] **Y.N. Chuang**, H. Zhou, P. Sarma, P. Gopalan, J. Boccio, S. Bolouki, and X. Hu. "Learning to Route with Confidence Tokens" *International Conference on Machine Learning*
- [EMNLP' 24] **Y.N. Chuang***, G. Wang*, R. Tang, S. Zhong, J. Yuan, H. Jin, Z. Liu, V. Chaudhary, S. Xu, J. Caverlee, and X. Hu. "Taylor Unswift: Secured Weight Release for Large Language Models via Taylor Expansion" *Annual Conference of the North American Chapter of the ACL*
- [NAACL' 24] **Y.N. Chuang**, T. Xing, C.Y. Chang, Z. Liu, X. Chen, and X. Hu. "Learning to Compress Prompt in Natural Language Formats" *Annual Conference of the North American Chapter of the ACL*
- [NAACL' 24 Finding] **Y.N. Chuang***, R. Tang*, and X. Hu. "Secure Your Model: A Simple but Effective Key Prompt Protection Mechanism for Large Language Models" *Finding of Annual Conference of the North American Chapter of the ACL*
- [ICLR' 23] **Y.N. Chuang***, G. Wang*, F. Yang, Q. Zhou, P. Tripathi, X. Cai and X. Hu. "CoRTX: Contrastive Learning for Real-time Explanations" *International Conference on Learning Representations*
- [ICML' 22 Spotlight] **Y.N. Chuang***, G. Wang*, M. Du, F. Yang, Q. Zhou, P. Tripathi, X. Cai and X. Hu. "Accelerating Shapley Explanation via Contributive Cooperator Selection" *International Conference on Machine Learning*
- [JBI] **Y.N. Chuang**, R. Tang, X. Jiang, and X. Hu. "SPeC: A Soft Prompt-Based Calibration on Performance Variability of Large Language Model in Clinical Notes Summarization" *Journal of Biomedical Informatics*
- [TKDD] **Y.N. Chuang**, K.H. Lai, R. Tang, M. Du, C.Y. Chang, N. Zou, and X. Hu. "Mitigating Relational Bias on Knowledge Graphs" *ACM Transactions on Knowledge Discovery from Data*
- [CIKM' 23] **Y.N. Chuang**, G. Wang et al., and X. Hu "DiscoverPath: A Knowledge Refinement and Retrieval System for Interdisciplinarity on Biomedical Research" *ACM International Conference on Information and Knowledge Management (CIKM'23 Best Demo Paper Honorable Mention)*
- [CIKM' 20] **Y.N. Chuang***, C.M. Chen*, C.J. Wang, M.F. Tsai, Y. Fang, and E.P. Lim. "TPR: Text-aware Preference Ranking for Recommender Systems" *ACM International Conference on Information and Knowledge Management*
- [UAI' 20] **Y.N. Chuang***, C.J. Wang, C.M. Chen, and M.F. Tsai. "Skewness Ranking Optimization for Personalized Recommendation" *Conference on Uncertainty in Artificial Intelligence (Oral)*
- [CACM] R. Tang, **Y.N. Chuang**, and X. Hu. "The Science of LLM-generated Text Detection" *The Communications of the ACM (CACM April Cover)*
- [TMLR] Y. Sui, **Y.N. Chuang**, G. Wang, et. al., and X. Hu. "Stop Overthinking: A Survey on Efficient Reasoning for Large Language Models" *Transactions on Machine Learning Research*
- [KDD' 25] C.Y. Chang, **Y.N. Chuang**, Z. Jiang, K.H. Lai, A. Jiang, N. Zou. "CODA: Temporal Domain Generalization via Concept Drift Simulator" *International Conference on Knowledge Discovery and Data Mining*
- [ACL' 25] M. Zhong, G. Wang, **Y.N. Chuang**, N. Zou. "Quantized Can Still Be Calibrated: A Unified Framework to Calibration in Quantized Large Language Models" *Annual Meeting of the Association for Computational Linguistics*
- [ACL' 25 Finding] J. Zhang, J. Yuan, A. Wen, H. Le, **Y.N. Chuang**, S. Choi, R. Chen, X. Hu. "ReasonerRank: Redefining Language Model Evaluation with Ground-Truth-Free Ranking Frameworks" *Finding of Annual Meeting of the Association for Computational Linguistics*
- [EMNLP' 25 Finding] L. Zhang, **Y.N. Chuang**, G. Wang, R. Tang, X. Cai, R. Shenoy, and X. Hu "A Decoupled Multi-Agent Framework for Complex Text Style Transfer"

[EMNLP' 25 Finding] Z. Xu, G. Wang, G. Zheng, **Y.N. Chuang**, A. Szalay, X. Hu, and V. Braverman "Self-Ensemble: Mitigating Confidence Distortion for Large Language Models"

[NAACL' 25 Finding] Y. Wang*, J. Yuan*, **Y.N. Chuang**, et al, and X. Hu, "DHP Benchmark: Are LLMs Good NLG Evaluators?" *Finding of Annual Conference of the North American Chapter of the ACL*

[ICML' 24] G. Wang, **Y.N. Chuang**, F. Yang, M. Du, C.Y. Chang, et al., and X. Cai, and X. Hu. "TVE: Learning Meta-attribution for Transferable Vision Explainer" *International Conference on Machine Learning*

[EMNLP' 24 Finding] J. Yuan*, H. Liu*, S. Zhong*, **Y.N. Chuang**, et al., and X. Hu. "KV Cache Compression, But What Must We Give in Return? A Comprehensive Benchmark of Long Context Capable Approaches" *Finding of Empirical Methods in Natural Language Processing*

Preprints and Under Review

[Submitted NeurIPS' 25] **Y.N. Chuang***, F. Lou*, G. Wang, H. Le, et. al., V. Braverman, V. Chaudhary, and X. Hu. "AutoL2S: Auto Long-Short Reasoning for Efficient Large Language Models"

[Submitted NeurIPS' 25] **Y.N. Chuang***, L. Yu, G. Wang, et. al., V. Braverman, and X. Hu. "Confident or Seek Stronger: Exploring Uncertainty-Based On-device LLM Routing From Benchmarking to Generalization"

[Submitted NeurIPS' 25] H. Le, S. Zhong, Y. Lu, Y. Dou, J. Yuan, **Y.N. Chuang**, et. al., X. Hu. "FAFO: Lossless KV Cache Compression with Draftless Fumble Decoding"

[Submitted TMLR] **Y.N. Chuang**, G. Wang, C.Y. Chang, R. Tang, S. Zhong, F. Yang, M. Du, X. Cai, V. Braverman, and X. Hu "Towards Faithful Explanations for Large Language Models"

[Submitted TMLR] G. Wang, **Y.N. Chuang**, H. Chen, Y. Chen, Z. Jiang, M. Bendre, M. Das, Z. Liu, J. Yuan, and X. Hu. "LEMO: Learning Shapley Manifold for Faithful Explanation"

[Submitted TMLR] **Y.N. Chuang**, G. Wang, F. Yang, Z. Liu, X. Cai, M. Du, and X. Hu. "Efficient XAI Techniques: A Taxonomic Survey"

OPEN SOURCE PACKAGE

DiscoverPath: A Knowledge Refinement and Retrieval System for Interdisciplinarity on Biomedical Research (CIKM'23 Best Demo Paper Honorable Mention)

- *Project Leader.* Designed a KG-based retrieval system designed for biomedical research that aims to assist biomedical researchers in dynamically refining their queries and effectively retrieving articles.

LTSM-bundle: Large Time Series Models Training and Benchmark Library

- *Project Leader.* Designed the package architectures with CI/CD pipeline for large-scale time series data.

SMORe: Modularize Graph Embedding for Recommendation

- *Developer.* Constructed a large-scale network embedding library for recommendation systems on online streaming services which was developed under C++ with multi-thread processing techniques

HONORS AND AWARDS

- Rice D2K Fellowship, Rice University	<i>Sep. 2025</i>
- Study Abroad Fellowship, Ministry of Education, Taiwan	<i>May. 2025</i>
- Ken Kennedy Institute Fellowship, Rice University	<i>Nov. 2024</i>
- Doctoral Forum Travel Award, SDM' 24	<i>Mar. 2024</i>
- CIKM 2023 Best Demo Paper Honorable Mention	<i>Oct. 2023</i>
- 4th Place at ACM RecSys Challenge	<i>Sep. 2020</i>

PROFESSIONAL SERVICES

Reviewer (Since 2020): NeurIPS, ICLR, ICML, ACL, EMNLP, NAACL, AAAI, IJCAI, KDD, WSDM, CIKM, IEEE TPAMI, IEEE TAI, IEEE TIST, IEEE ICHI