

Xuan Phi Nguyen

🌐 [linkedin.com/in/xuanphinguyen](https://www.linkedin.com/in/xuanphinguyen) 🐙 [nxphi47.github.io](https://github.com/nxphi47) ✉ nxphi47@gmail.com 🎓 G-Scholar

EXPERIENCE

Salesforce AI Research

Senior Research Scientist

May 2024 – Present

- Built SFR-RL, a memory-efficient, high-throughput full-parameter training infrastructure for AI agents; supporting 120B–1T models at 1M context on minimally 16–96 GPUs. Its Mixture-of-Parallelisms pipeline achieved up to 8x faster than FSDP2 + 5D-parallelisms.
- Developed SFR-VibeTrain, an autonomous agent that builds and improves other AI agents. It does so by iteratively crafting and executing RL training workflows and/or auto-harness searches, until specified goals are achieved.
- Shipped SFR-DeepResearch, a single-agent system trained with reinforcement learning (RL) for multi-step research with external tools on complex reasoning tasks.
- Drove SFR-RAG, small LLMs for Retrieval Augmented Generation (RAG), outperforming GPT-4o in certain key RAG benchmarks; highlighted by CEO Marc Benioff and covered in the press.

Alibaba Group

Senior Algorithm Engineer (AI Research)

April 2023 – May 2024

- Led research and engineering for SeaLLMs, the first open large language models for Southeast Asian languages, surpassing GPT-3.5 in these languages; highlighted by regional media.

Facebook AI Research (FAIR), Meta

AI Research Intern

May 2022 – Sep 2022

- Improved speech-to-speech translation via synthetic data generation and augmentation; paper at ICASSP 2023.

Facebook AI Research (FAIR), Meta

AI Research Intern

May 2021 – Sep 2021

- Developed pseudo-parallel data mining and integration for unsupervised machine translation (no parallel human annotations); paper at ICLR 2022.

Salesforce AI Research

AI Research Intern

May 2019 – Aug 2019

- Developed a tree-structured attention mechanism for the Transformer, achieving state-of-the-art on machine translation benchmarks; paper at ICLR 2020.

Visa Inc.

Software Engineer Intern

May 2017 – Jul 2017 and Jan 2018 – Jul 2018

- Built a convolutional model for scripted text classification at 98% accuracy with $\sim 1000\times$ fewer parameters than the prior baseline.

EDUCATION

Nanyang Technological University

Ph.D., Computer Science & Artificial Intelligence

2023

- Thesis: “**Improving Neural Machine Translation: Data-Centric Approaches**”.

Nanyang Technological University

B.Eng., Electrical & Electronic Engineering (First Class Honours)

2019

- Dean’s List (top 5% of cohort by GPA), two academic years.

SELECTED RESEARCH PUBLICATIONS

- Xuan-Phi Nguyen, Shrey Pandit, Yiran Zhao, Semih Yavuz, Silvio Savarese, Shafiq Joty (2026). *Mixture-of-Parallelisms: Towards Memory-Efficient Training Stack for Mixture-of-Experts Models*. **System Card**.
- Xuan-Phi Nguyen, Shrey Pandit, Austin Xu, Caiming Xiong, Shafiq Joty (2026). *Least-Loaded Expert Parallelism: Load Balancing An Imbalanced Mixture-of-Experts*. **ICML 2026**.
- Xuan-Phi Nguyen, Shrey Pandit, Revanth Gangi Reddy, Austin Xu, Silvio Savarese, Caiming Xiong, Shafiq Joty (2025). *SFR-DeepResearch: Towards effective reinforcement learning for autonomously reasoning single agents*. **Technical Report**.
- Austin Xu, Xuan-Phi Nguyen, Yilun Zhou, Chien-Sheng Wu, Caiming Xiong, Shafiq Joty (2025). *Foundational Automatic Evaluators: Scaling Multi-Task Generative Evaluator Training for Reasoning-Centric Domains*. **ICLR 2026**.
- Shrey Pandit, Xuan-Phi Nguyen, Yifei Ming, Austin Xu, Jiayu Wang, Caiming Xiong, Shafiq Joty (2025). *Synthesizing Agentic Data for Web Agents with Progressive Difficulty Enhancement Mechanisms*. In submission.

- Shrey Pandit, Austin Xu, Xuan-Phi Nguyen, Yifei Ming, Caiming Xiong, Shafiq Joty (2025). *Hard2Verify: A Step-Level Verification Benchmark for Open-Ended Frontier Math*. **ACL 2026**.
- Yifei Ming, Zixuan Ke, Xuan-Phi Nguyen, Jiayu Wang, Shafiq Joty (2025). *Helpful Agent Meets Deceptive Judge: Understanding Vulnerabilities in Agentic Workflows*. In submission.
- Austin Xu, Yilun Zhou, Xuan-Phi Nguyen, Caiming Xiong, Shafiq Joty (2025). *J4R: Learning to Judge with Equivalent Initial State Group Relative Policy Optimization*. **ACL 2026**.
- Zixuan Ke, Austin Xu, Yifei Ming, Xuan-Phi Nguyen, Ryan Chin, Caiming Xiong, Shafiq Joty (2025). *MAS-ZERO: Designing Multi-Agent Systems with Zero Supervision*. **SEA@NeurIPS (Oral) 2025**.
- Revanth Gangi Reddy, Tarun Suresh, JaeHyeok Doo, Ye Liu, Xuan Phi Nguyen, Yingbo Zhou, Semih Yavuz, Caiming Xiong, Heng Ji, Shafiq Joty (2025). *SweRank: Software Issue Localization with Code Ranking*. **ICLR 2026**.
- Zixuan Ke, Yifei Ming, Xuan-Phi Nguyen, Caiming Xiong, Shafiq Joty (2025). *Demystifying domain-adaptive post-training for financial LLMs*. **EMNLP 2025 - Oral**.
- Hailin Chen, Fangkai Jiao, Mathieu Ravaut, Nawshad Farruque, Xuan Phi Nguyen, Chengwei Qin, Manan Dey, Bosheng Ding, Caiming Xiong, Shafiq Joty, Yingbo Zhou (2025). *StructTest: Benchmarking LLMs' Reasoning through Compositional Structured Outputs*. **CoRR 2024**.
- Yifei Ming, Senthil Purushwalkam, Shrey Pandit, Zixuan Ke, Xuan-Phi Nguyen, Caiming Xiong, Shafiq Joty (2025). *FaithEval: Can Your Language Model Stay Faithful to Context, Even If "The Moon is Made of Marshmallows"*. **ICLR 2025**.
- Zhenmei Shi, Yifei Ming, Xuan-Phi Nguyen, Yingyu Liang, Shafiq Joty (2024). *Discovering the gems in early layers: Accelerating long-context llms with 1000x input token reduction*. Preprint.
- Xuan-Phi Nguyen, Shrey Pandit, Senthil Purushwalkam, Austin Xu, Hailin Chen, Yifei Ming, Zixuan Ke, Silvio Savarese, Caiming Xiong & Shafiq Joty (2024). *SFR-RAG: Towards Contextually Faithful LLMs*. **Technical Report**.
- Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, Hang Zhang & Lidong Bing (2024). *SeaLLMs - Large Language Models for Southeast Asia*. In Proceedings of the Association for Computational Linguistics (**ACL 2024**), Demo Track.
- Xuan-Phi Nguyen, Sharifah Mahani Aljunied, Shafiq Joty & Lidong Bing (2024). *Democratizing LLMs for Low-Resource Languages by Leveraging their English Dominant Abilities with Linguistically-Diverse Prompts*. In Proceedings of the Association for Computational Linguistics (**ACL 2024**).
- Chenhui Shen, Liying Cheng, Xuan-Phi Nguyen, Yang You & Lidong Bing (2023). *Large Language Models are Not Yet Human-level Evaluators for Abstractive Summarization*. The 2023 Conference on Empirical Methods in Natural Language Processing (**EMNLP 2023**) - Findings.
- Chenhui Shen, Liying Cheng, Xuan-Phi Nguyen, Yang You & Lidong Bing (2023). *A Hierarchical Encoding-decoding Scheme for Abstractive Multi-document Summarization*. The 2023 Conference on Empirical Methods in Natural Language Processing (**EMNLP 2023**) - Findings.
- Xuan-Phi Nguyen, Sravya Popuri, Chaghan Wang, Yun Tang, Ilia Kulikov & Hongyu Gong (2023). *Improving Speech-to-Speech Translation Through Unlabeled Text*. 2023 IEEE International Conference on Acoustics, Speech, and Signal Processing (**ICASSP 2023**).
- Xuan-Phi Nguyen, Shafiq Joty, Wu Kui & Aw Ai Ti (2022). *Refining Low-Resource Unsupervised Translation by Language Disentanglement of Multilingual Model*. 36th Conference on Neural Information Processing Systems (**NeurIPS 2022**).
- Xuan-Phi Nguyen, Hongyu Gong, Yun Tang, Chaghan Wang, Philipp Koehn & Shafiq Joty (2022). *Contrastive Clustering to Mine Pseudo Parallel Data for Unsupervised Translation*. In International Conference on Learning Representations (**ICLR 2022**).
- Xuan-Phi Nguyen, Shafiq Joty, Thanh-Tung Nguyen, Wu Kui, & Ai Ti Aw (2021). *Cross-model Back-translated Distillation for Unsupervised Machine Translation*. In Proceedings of the 38th International Conference on Machine Learning (**ICML 2021**).
- Xuan-Phi Nguyen, Shafiq Joty, Wu Kui & Aw Ai Ti (2020). *Data Diversification: A Simple Strategy For Neural Machine Translation*. 34th Conference on Neural Information Processing Systems (**NeurIPS 2020**).
- Xuan-Phi Nguyen, Shafiq Joty, Steven Hoi & Richard Socher (2020). *Tree-Structured Attention with Hierarchical Accumulation*. In International Conference on Learning Representations (**ICLR 2020**).
- Thanh-Tung Nguyen, Xuan-Phi Nguyen, Shafiq Joty & Xiaoli Li (2021). *A Conditional Splitting Framework for Efficient Constituency Parsing*. In Proceedings of Association for Computational Linguistics (**ACL 2021**).
- Thanh-Tung Nguyen, Xuan-Phi Nguyen, Shafiq Joty & Xiaoli Li (2021). *RST Parsing from Scratch*. In the North American Chapter of the Association for Computational Linguistics (**NAACL 2021**).
- Xuan-Phi Nguyen*, Thanh-Tung Nguyen*, Shafiq Joty & Xiaoli Li (2020). *Differentiable Window for Dynamic Local Attention*. In Proceedings of the Association for Computational Linguistics (**ACL 2020**). Co-

first-authored.

- Thanh-Tung Nguyen, Xuan-Phi Nguyen, Shafiq Joty & Xiaoli Li (2020). *Efficient Constituency Parsing by Pointing*. In Proceedings of the Association for Computational Linguistics (**ACL 2020**).
- Xuan-Phi Nguyen & Shafiq Joty (2019). *Phrase-based Attentions*. arXiv preprint arXiv:1810.03444.
- P. X. Nguyen, Z. Lu, W. Huang, S. Huang, A. Katsuki & Z. Lin (2019). *Medical image segmentation with stochastic aggregated loss in a unified U-Net*. In 2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI) (**IEEE BHI 2019**), Chicago, USA.

HONORS & AWARDS

Reviewer for major AI conferences

- NeurIPS 2021-2025; ICLR 2022-2026; ICML 2023-2026, ACL 2021-2026.

2021 Singapore Data Science Consortium (SDSC) Dissertation Research Award

Jan - Dec 2022

- Awarded for innovative and impactful thesis work of PhD students whose research interests are in data science.

NeurIPS Student Scholar Award

Nov 2022

- Awarded for PhD students with first-authored papers at NeurIPS 2022 conference at New Orleans, USA.

A*STAR Computing and Information Science (ACIS) Scholarship

Jul 2019 – Jul 2023

- Fully funded by A*STAR to pursue the doctorate program in NTU, Singapore.

The Global Undergraduate Awards - Highly Commended Paper

Sep 2019

- Top 10% out of more than 3500 papers submitted globally.

Shopee - IET Machine Learning Challenge Competition - 2nd Runner-up

Mar 2018

- Designed ensemble techniques to achieve 2nd Runner-up prize with 85.662% accuracy in image classification task.

SUTD What The Hack Hackathon - Best Software Hack - Best Security Hack

Sep 2017

- Eye Gesture Control Application: VR device powered by neural networks to help the disabled to perform daily tasks.

Singapore Scholarship

Jul 2015 – Jul 2019

- Fully funded by Ministry of Foreign Affairs of Singapore to pursue the undergraduate program in NTU, Singapore.