Chien Nguyen

chiennv.hust@gmail.com Google Scholar chiennv2000.github.io
linkedin.com/in/chiennv2000 github.com/chiennv2000

Research Summary

My research focuses on efficient and scalable generative modeling, with an emphasis on designing novel architectures and training strategies for Large Language Models (LLMs). I am particularly interested in pushing the boundaries of efficient inference, long-context modeling, and reasoning.

Education

University of Oregon, USA *Sep 2023 – Sep 2027*
PhD in Computer Science, advised by Prof. Thien Huu Nguyen *(Expected)*

Hanoi University of Science and Technology, Vietnam *Sep 2018 – May 2023*
BSc in Computer Science

Experience

Google DeepMind *Mountain View, USA*
Student Researcher *Sep 2025 – Jan 2026*

- Improved LLM inference efficiency by enabling cross-model KV-cache compatibility via KV-cache fusion.

Adobe Research *San Jose, USA*
Research Scientist Intern *Jun 2024 – Sep 2025*

- Developed efficient architectures and training methods for long-context LLMs.
- Developed small multilingual language models under limited-data constraints.

VinAI Research (now Qualcomm AI) *Hanoi, Vietnam*
AI Research Resident *Mar 2022 – Sep 2023*

- Research Topics: Structured Information Extraction, Multilingual LLMs.

Selected Publications and Preprints

2026 Chien Van Nguyen, Chaitra Hegde, Van-Cuong Pham, Ryan A. Rossi, Franck Dernoncourt, Thien Huu Nguyen. **Orthrus: Memory-Efficient Parallel Token Generation via Dual-View Diffusion.** *arXiv preprint* `arXiv:2605.12825`.

2026 Chien Van Nguyen, Ryan A. Rossi, Linh Ngo, Franck Dernoncourt, Thien Huu Nguyen. **Octopus: Gated Selective Attention for Memory-Bounded Long-Context Inference in Large Language Models.** In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2026.

2026 Chien Van Nguyen, Huy Nguyen, Ryan A. Rossi, Trung Bui, Nikos Vlassis, Franck Dernoncourt, Thien Huu Nguyen. **Lizard: An Efficient Linearization Framework for Large Language Models.** In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2026.

2025 Chien Van Nguyen, Huy Huu Nguyen, Ryan A. Rossi, Trung Bui, Viet Lai, Franck Dernoncourt, Thien Nguyen. **Taipan: Efficient and Expressive State Space Language Models with Selective Attention.** *arXiv preprint* `arXiv:2410.18572`.

2024 Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, Thien Huu Nguyen. **CulturaX: A Cleaned, Enormous, and Multilingual Dataset for LLMs in 167 Languages.** In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (COLING-LREC)*, 2024.

2024 Hieu Man, Chien Van Nguyen, Nghia Trung Ngo, Linh Ngo, Franck Dernoncourt, Thien Huu Nguyen. **Hierarchical Selection of Important Context for Generative Event Causality Identification with Optimal Transports.** In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (COLING-LREC)*, 2024.

2023 Viet Lai*, Chien Van Nguyen*, Nghia Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A. Rossi, Thien Huu Nguyen. **Okapi: Instruction-tuned Large Language Models in Multiple Languages with Reinforcement Learning from Human Feedback.** In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP)*, 2023.

2023 Chien Van Nguyen, Huy Huu Nguyen, Franck Dernoncourt, Thien Huu Nguyen. **Transitioning Representations between Languages for Cross-lingual Event Detection via Langevin Dynamics.** In *Findings of the Association for Computational Linguistics: EMNLP 2023 (ACL)*, 2023

2023 Chien Van Nguyen, Linh Ngo Van, Thien Huu Nguyen. **Retrieving Context to Align Representations for Cross-lingual Event Detection.** In *Findings of the Association for Computational Linguistics: ACL 2023 (ACL)*, 2023

2023 Chien Van Nguyen, Hieu Man Duc Trong, Thien Huu Nguyen. **Contextualized Soft Prompts for Extraction of Event Arguments.** In *Findings of the Association for Computational Linguistics: ACL 2023 (ACL)*, 2023

2023 Huy Nguyen, Chien Van Nguyen, Linh Ngo, Luu Anh Tuan, Thien Huu Nguyen. **A Spectral Viewpoint on Continual Relation Extraction.** In *Findings of the Association for Computational Linguistics: EMNLP 2023 (EMNLP)*, 2023

*Equal contribution.

Projects

Orthrus: A Dual-View Architecture github.com/chiennv2000/orthrus *Lead Author*

- Designed a dual-view framework that augments a *frozen* autoregressive LLM with a lightweight diffusion head, enabling parallel token generation – achieving up to 7.8× throughput speedup with $\mathcal{O}(1)$ memory overhead.

GPT-OSS Efficient Finetuning nlp-uoregon/oregon_gpt_oss_patching *Lead Author*

- Implemented custom Triton-based flash attention kernels with a complete backward pass for OpenAI's GPT-OSS models (20B and 120B), unlocking memory-efficient full finetuning on long contexts on a single node with a 2.5× throughput gain.

Vistral-7B-Chat Viet-Mistral/Vistral-7B-Chat *Lead Author*

- Built a SOTA Vietnamese conversational LLM; outperforms ChatGPT and Gemini on Vietnamese benchmarks at release; **200,000+ downloads** on Hugging Face.

Technologies

Languages & Frameworks:	C/C++, PyTorch, JAX/Flax, CUDA, Triton, Cutlass CuTe, CuTeDSL, cuTile
Libraries:	HuggingFace Ecosystem, vLLM, SGLang
Tools:	Git, Bash, AWS, SLURM, RunAI

Academic Service

Conference Reviewer: ACL, EMNLP, AAAI, ICLR, NeurIPS

Awards

Lokey Award (2023–24)
Prestigious award from the Department of Computer Science, University of Oregon.

First Prize – *BKAI–NAVER Challenge 2022*
Ranked 1st among 80 teams in Intent Detection and Slot Filling.

First Prize – *SoICT–IBM Hackathon 2020*
Ranked 1st out of 20 finalist projects for developing an open-domain chatbot.