My (Chiffon) Nguyen

San Francisco, CA, USA | hi@mychiffonn.com | github.com/mychiffonn | mychiffonn.com

Research Interests

Current and future AI systems aligned with diverse human experiences across social contexts: **AI alignment** (pluralistic alignment, generalization, mechanistic interpretability), **multilingual and multicultural AI**, **socially-aware AI**, **human-AI collaboration**. Keywords: Natural Language Processing, Machine Learning, Human-AI Interaction.

Education

Minerva University, College of Computational Sciences Sep 2021 – May 2025
B.Sc in Computational Sciences (Machine Learning and Statistics), GPA: 3.7/4.0 *San Francisco, CA, USA*

- **Relevant Coursework**: Machine Learning (A), AI Ethics, Bayesian Modeling (A), Statistical Modeling and Causal Inference (A), Optimization Methods (A), Probability and Statistics (A-), Software Engineering
- **Global Experience**: Seoul (South Korea), Taipei, Hyderabad (India), Buenos Aires (Argentina), Berlin (Germany)

Selected Publication

P=Preprint, **S**=In Submission, **J**=Journal, **C**=Conference, **W**=Workshop. Full publication list on Google Scholar.

[S2] **LINKS: Linguistically-Informed Mnemonics for Monolingual and Cross-Lingual Vocabulary Acquisition**
My Chiffon Nguyen
In submission for ACL Building Educational Applications (BEA) Workshop, 2026.

[S1] **Anthropogenic Regional Adaptation in Multimodal Vision-Language Model**
Samuel Cahyawijaya, Peerat Limkonchotiwat, Tack Hwa Wong, Carlos Rafael Catalan, Manuel Antonio Rufino, Hitesh Laxmichand Patel, Amit Agarwal, Muhammad Reza Qorib, Vicky Feliren, Holy Lovenia, Aye Hninn Khine, Frederikus Hudi, David Anugraha, Romrawin Chumpu, Alham Fikri Aji, Viet Thanh Pham, Minghan Wang, Mohamed Fazli Mohamed Imam, Ruochen Zhang, Joseph Marvin Imperial, Khumaisa Nur'aini, Do Xuan Long, Musa Izzanardi Wijanarko, Joel Ruben Antony Moniz, Patrick Amadeus Irawan, Hanif Muhammad Zhafran, Salsabila Zahirah Pranida, Isaiah Edri W. Flores, Jun Kevin, Jostin Jerico Rosal, Patricia Nicole Monderin, Kun Kerdthaisong, Ahmad Mustafid, **My Chiffon Nguyen**, and 12 others
In submission for European Conference on Computer Vision (ECCV), 2026.

[P1] **CommonLID: Re-evaluating State-of-the-Art Language Identification Performance on Web Data**
Pedro Ortiz Suarez, Laurie Burchell, Catherine Arnett, Rafael Mosquera-Gómez, Sara Hincapie-Monsalve, Thom Vaughan, Damian Stewart, Malte Ostendorff, Idris Abdulmumin, Vukosi Marivate, Shamsuddeen Hassan Muhammad, Atnafu Lambebo Tonja, Hend Al-Khalifa, Nadia Ghezaiel Hammouda, Verrah Otiende, Tack Hwa Wong, Jakhongir Saydaliev, Melika Nobakhtian, Muhammad Ravi Shulthan Habibi, Chalamalasetti Kranti, Carol Muchemi, Khang Nguyen, Faisal Muhammad Adam, Luis Frentzen Salim, Reem Alqifari, Cynthia Amol, Joseph Marvin Imperial, Ilker Kesen, Ahmad Mustafid, Pavel Stepachev, Leshem Choshen, David Anugraha, Hamada Nayel, Seid Muhie Yimam, Vallerie Alexandra Putra, **My Chiffon Nguyen**, and 61 others
Preprint, 2026.

Research Experience

AI Research Fellow (Multilingual Agentic Evaluation) Feb 2026 – Expected Jun 2026
SEACrowd, SEACrowd 2026 Research Apprenticeship *Remote*

- Evaluating conversational agents across multilingual and cross-lingual settings by extending Tau2-Bench
- Designing a visually- and culturally-grounded food delivery domain for agentic evaluation, incorporating region-specific menus, interaction patterns, and code-switching user queries
- Mentors: Dr. Samuel Cahyawijaya (Cohere Labs) and Patomporn Payoungkhamdee (VISTEC Thailand)

AI Research Mentee (Chain-of-thought Monitorability; Project Lead) Oct 2025 – Expected Apr 2026
Algoverse AI Research, AI Research Program Fall 2025 (Mentor: Yeonwoo Jang) *Remote*

- Investigating whether weaker AI models can detect when stronger models deliberately underperform on safety-critical evaluations, measuring monitoring success across 20 pairs of open-weight models (Qwen3, Olmo3, GPT-OSS) [Report]
- Designed end-to-end monitoring pipeline with `Inspect AI` and `Inspect Scout`, and estimating capability gaps via the Epoch.ai Capabilities Index to systematically select monitor-target model pairs

Collaborator (Mapping Multilingual Representations with Sparse Autoencoders) Mar 2026 (expected)
Cohere Tiny Aya Expedition (Mentor: Dr. Tom Hosking) *Remote*

- Applying Sparse Autoencoders (SAEs) to Tiny Aya (3.3B, 70+ languages) to identify universal vs. language-specific internal features, in a team of four [Code]

- Building multilingual data pipeline over Flores-200 and CulturaX for feature comparison across 70 languages
- Designing evaluation pipeline measuring generation quality of Tiny-Aya models before and after feature steering, using LaBSE semantic similarity, script conformity, and LLM-as-a-judge with Cohere's Command A models

Selected Projects

Replication: Unsupervised Elicitation of Language Models (github.com/mychiffonn/icm) Dec 2025

Replicated Wen et al. (2025)'s Internal Coherence Maximization (ICM), which elicits human-interpretable concepts from base language models by maximizing mutual predictability and local consistency among concept-related examples.

Mini-LLaMA2 PyTorch Implementation (github.com/mychiffonn/cmu-advanced-nlp-minllama) May 2024

- Implemented the core architecture of Llama-2 from scratch in PyTorch, including Rotary Positional Embeddings (RoPE), RMSNorm, and SwiGLU activation functions
- Developed a custom training loop with AdamW optimization to pretrain on TinyStories and fine-tune for sentiment classification (SST-5), achieving coherent text generation

Replication: Synthetic Control (Causal Inference) (github.com/mychiffonn/synthetic-control-rep) Dec 2023

- Replicated Chrisinger (2021)'s synthetic control analysis of Philadelphia's SNAP benefit redemption in R, analyzing policy impacts across 4 counties and 50+ months of longitudinal data
- Extended the research by conducting novel leave-one-out robustness analysis, revealing donor pool sensitivity and model instability in the original findings

Teaching & Mentoring Experience

Curious Cardinals, *AI & Data Science*, Mentor Nov 2025 – Present

- Mentoring two high school science fair projects: (1) Association between HEMA genes and Parkinson disease and (2) Robustness of Fact-Checking Language Models under Evidence Corruption and Language Shifts

Minerva University, *PR51 Programming with Python*, Lead Peer Tutor and Data Analyst Spring 2025

- Taught 40+ first-year students from 20+ countries in **weekly hands-on programming labs** for 11 weeks, covering Python, OOP, debugging, security, and computing fundamentals
- Analyzed student performance data and tutor surveys across 11 weeks to identify 12 learning bottlenecks, improving engagement metrics by 15% for the next cohort

Minerva University, *FA50/FA51 Logic, Probability & Statistics*, Lead Teaching Assistant Fall 2023 – Spring 2024

- Guided **150+ students each semester** across four semesters in formal logic, probability, statistics, algorithmic thinking, and simulation, through weekly office hours
- Provided **formative assessment on 25 quizzes** for 50 students to correct and shape their learning
- Assisted professors in **grading** three math and programming assignments per semester

Leadership & Services

SEACrowd **Communications Associate & Web Design Engineer** Aug 2025 – Present

Developh Vietnam **Head of Public Relations** Dec 2019 – Dec 2021

Skills

- **Programming Languages**: Python, TypeScript, SQL, R, Bash
- **Machine Learning**: PyTorch, Inspect AI, unsloth, trl, scikit-learn, LangGraph, LlamaIndex
- **Web/App Development**: Astro, React, FastAPI, Flask, Express.js, PostgreSQL, TailwindCSS, shadcn/ui
- **Tools & DevOps**: Git, Docker, Python tooling (uv, ruff, ty), Render, Netlify, LaTeX, Zotero, Typst
- **Languages**: Vietnamese (native), English (fluent/C1), Mandarin Chinese (lower-intermediate/HSK 4)

Certificates

- **Advanced Web Development**, CodePath (drive.google.com/file/d/1n4dHj4TFM8HWlDXMTt9ZGjEXVIpkP-F-)
- **Natural Language Specialization**, deeplearning.ai (coursera.org/verify/specialization/3FJ3W7QJX8GK) Nov 2023
- **Applied Data Science**, World Quant University (credly.com/badges/2e1e6902-aae4-47c4-97e2-0ad9265e5561)Aug 2023
- **Machine Learning Specialization**, deeplearning.ai (coursera.org/verify/specialization/G9898XKB9EAV) Jun 2022

Last Updated: Mar 23, 2026