My (Chiffon) Nguyen

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

Research Interests

Current and future AI that are socially responsible and beneficial for diverse groups of humans and social contexts

- **AI alignment** (pluralistic alignment, OOD generalization, bias & fairness, trustworthy AI) & **AI control**
- **multilingual and multicultural AI**, focusing on language, speech, and vision-language models
- **human-AI collaboration & socially aware AI** grounded in social science, Human-Computer Interaction research

Education

Minerva University, College of Computational Sciences Sep 2021 – May 2025

B.Sc in Computational Sciences (Machine Learning and Statistics), GPA: 3.67/4.00 *San Francisco, CA, USA*

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

Publication

P=Preprint, **R**=Under Review, **J**=Journal, **C**=Conference, **W**=Workshop. Full publication list on Google Scholar.

[R1] **Anthropogenic Regional Adaptation in Multimodal Vision-Language Model**
Samuel Cahyawijaya, Peerat Limkonchotiwat, Tack Hwa Wong, et al. (including **My Chiffon Nguyen**)
Under review at European Conference on Computer Vision (ECCV), 2026.

[C1] **CommonLID: Re-evaluating State-of-the-Art Language Identification Performance on Web Data**
Pedro Ortiz Suarez, Laurie Burchell, Catherine Arnett, et al. (including **My Chiffon Nguyen**)
Association of Computational Linguistics (ACL) Main Conference, 2026.

[J1] **A benchmark of expert-level academic questions to assess AI capabilities**
Center for Al Safety, Scale AI, HLE Contributors Consortium
Nature, 2026.

Research Experience

AI Research Mentee (Multilingual Agentic Evaluation) Feb – Expected Jun 2026

SEACrowd, SEACrowd 2026 Research Apprenticeship *Remote*

- Evaluating conversational agents across five multilingual, cross-lingual, and localized settings by extending Tau3-Bench
- Mentors: Dr. Samuel Cahyawijaya (Cohere Labs) and Patomporn Payoungkhamdee (VISTEC Thailand)
- Target EMNLP 2026 main conference submission

AI Research Lead Mentee (Chain-of-thought Monitoring & Evaluation) Oct 2025 – Expected Apr 2026

Algoverse AI Research, AI Research Program Fall 2025 (Mentor: Yeonwoo Jang (MATS 8.0)) *Remote*

- Investigating whether weaker AI models can detect sandbagging (deliberate underperformance) in stronger models on safety-critical evaluations [Report]
- Designed end-to-end monitoring pipeline (Inspect AI + Inspect Scout) and evaluated monitoring success rates across 20 pairs of open-weight reasoning models (Qwen3, Olmo3)

Collaborator (Multilingual Representations with Sparse Autoencoders) Mar 2026

Cohere Tiny Aya Expedition (Mentor: Dr. Tom Hosking) *Remote*

- Applied Sparse Autoencoders (SAEs) to multilingual LLM (Tiny Aya, 3.3B, 70+ languages) to identify universal vs. language-specific internal features [Code] [HuggingFace]
- Evaluated generation quality of Tiny-Aya models before and after feature steering across 67 languages, using LaBSE semantic similarity, script conformity, and LLM-as-a-judge, comparing with Flores-200 and CulturaX datasets

Machine Learning Research Assistant Jun – Aug 2024

Landshut University of Applied Sciences, AI & Mixed Reality Lab *Landshut, Bavaria, Germany*

- Supervisors: Prof. Eduard Kromer and Prof. Sandra Eisenreich
- Topic: 3D object detection in point clouds, point cloud registration

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

Selected Projects

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: 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.

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 analysis with leave-one-out robustness analysis, revealing original donor pool sensitivity and model instability

Leadership & Volunteering

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

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

Technical Skills

- **Programming Languages**: Python, SQL, Bash, TypeScript, R
- **Machine Learning**: PyTorch, scikit-learn, unsloth, trl, Inspect AI, LangGraph, LlamaIndex
- **Web/App Development**: Astro, React, FastAPI, Express.js, PostgreSQL, TailwindCSS, shadcn/ui
- **Tools & DevOps**: Git, Docker, Python tooling (uv, ruff, ty), LaTeX, Zotero

Languages

Vietnamese (native), English (fluent/C1), Mandarin Chinese (lower-intermediate/B1/HSK 4)

Certificates

- **Tiny Aya Expedition**, Cohere Labs (credsverse.com/credentials/0e262642-a662-4d2b-ac4e-8ca35fc2f80c) Mar 2026
- **Advanced Web Development**, CodePath (drive.google.com/file/d/1n4dHj4TFM8HWlDXMTt9ZGjEXVIpkP-F-)
- **Natural Language Specialization**, deeplearning.ai (coursera.org/verify/specialization/3FJ3W7QJX8GK) Nov 2023
- **Machine Learning Specialization**, deeplearning.ai (coursera.org/verify/specialization/G9898XKB9EAV) Jun 2022

Last Updated: Apr 07, 2026