Tra My (Chiffon) Nguyen

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

Research Interests

Current and future AI systems aligned with and representative of diverse human experiences: **AI safety** (pluralistic alignment, evaluation, cooperative AI, AI control), **multilingual and multicultural AI**, and **socially responsible AI**.

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 Rotation**: Seoul (South Korea), Chinese Taipei, Hyderabad (India), Buenos Aires (Argentina), Berlin (Germany)
- **Self-study**: Technical AI Alignment (ARENA, 2025), Introduction to AI Alignment (Bluedot Impact, 2025)

Research Experience

AI Research Mentee (Multilingual Agentic Evaluation) Feb 2026 — Present
SEACrowd, SEACrowd 2026 Research Apprenticeship *Remote*

- Project: Extending Tau2 Bench to low-resource languages, new cultural domain, and visual modality
- Mentors: Samuel Cahyawijaya (Cohere Labs, HKUST PhD) and Patomporn Payoungkhamdee (VISTEC PhD Student)

AI Research Fellow (Chain-of-thought Monitorability) Oct 2025 — Present
Algoverse AI Research, AI Research Program Fall 2025 (Mentor: Yeonwoo Jang (MATS 8 Scholar)) *Remote / US*

- Project: Chain-of-thought monitorability in relation to target-monitor capability gaps. [Code]
- Description: Evaluating relationships between success rates (ROC-AUC) of monitoring **sandbagging** and capability gap of monitor-target pairs within Qwen3 model family (8B–480B) (and cross-family), using **Inspect AI**

Qualitative Researcher & Data Scientist Intern Feb 2023 — Apr 2023
Human Resources Department, T-Hub (Telangana's innovation hub) *Hyderabad, Telegana, India*

- Designed preliminary mental health framework serving 1,500 employees across 650+ startups by analyzing 30+ hours of interviews and employee databases in Excel, identifying 5 key stress factors to inform T-Hub's pilot program
- Led primary research in a team of four, conducting 30 structured interviews and distributing surveys to 50 employees, generating actionable insights that shaped T-Hub's mental health strategy for India's largest startup ecosystem

Teaching & Mentoring Experience

Curious Cardinals, *Passion Project*, Mentor Nov 2025 — Present

- Computational neuroscience: Association between HEMA genes and Parkinson disease
- CCIR project: Robustness of Fact-Checking Language Models under Evidence Corruption

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
- Extracted **40 data-driven pedagogical insights** using Google Drive API, Google Sheet trackers, student and tutor surveys, improving hands-on learning and student engagement for the next class iteration by 15%

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

- Guided **150+ students each semester** for four semesters in formal logic, probability and 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

SEACrowd **Communications Assosicate & Design Engineer** Aug 2025 — Present

Selected Projects

More projects on mychiffonn.com/projects and github.com/mychiffonn

Replication: Unsupervised Elicitation of Language Models (github.com/mychiffonn/icm) Dec 2025
Replicated Wen et. al (2025)'s Internal Coherence Maximization, which elicits human concepts from base language models by maximizing mutual predictability and local consistency among concept-related examples.

SportConnect: Connect Through Local Sport Events (github.com/mychiffonn/sport-connect) Nov 2025
Scalable web application connecting users to local recreational sports events, featuring secure multi-provider OAuth authentication, comprehensive event management, easy event discovery and filtering, and real-time RSVP tracking. Two-person full-stack project with TypeScript, React, Express, PostgreSQL, TailwindCSS, DaisyUI, and BetterAuth.

Academic Portfolio Theme for Researchers (github.com/mychiffonn/website) Aug — Oct 2025
High-performance academic theme, enabling fast, accessible, and multilingual publishing of publications and technical blogs while ensuring top-tier SEO (99) and Lighthouse (100) scores using Astro, TailwindCSS, and shadcn/ui

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 critical components including Rotary Positional Embeddings (RoPE), RMSNorm, and SwiGLU activation functions
- Developed a custom training loop with AdamW optimization to pretrain the model on a small corpus and fine-tune it for sentiment classification tasks (SST-5), resulting in convergence and coherent text generation

Replication: Synthetic Control (Causal Inference) (github.com/mychiffonn/synthetic-control-rep) Dec 2023
- Replicated and extended Chrisinger (2021)'s analysis of Philadelphia's SNAP benefit redemption in R, analyzing policy impacts across 4 counties and 50+ months of longitudinal data; identified critical limitations in dataset reliability and magnitude discrepancies between original and replicated results
- Conducted new leave-one-out robustness analysis on synthetic control models, showing model instability

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

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
- **Languages**: Vietnamese (native), English (fluent/C2), Mandarin Chinese (lower-intermediate/band 4)

Last Updated: Feb 19, 2026