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, scalable oversight, safety evaluation), multilingual and multicultural AI, socially responsible AI, and AI ethics.

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 Collaborator Nov 2025 — Present
Cohere Labs Community (Advisor: Ilia Badanin) *Remote*

- Topic: Multilingual token optimization via cross-lingual embedding alignment

AI Safety Research Mentee Oct 2025 — Present
Algoverse AI Research (Advisor: Yeonwoo Jang) *Remote*

- Topic: Scaling behavior of chain-of-thought monitoring in white-box models / Code

Machine Learning Research Assistant Jun 2024 — Aug 2024
AI & Mixed Reality Lab, Landshut University of Applied Sciences *Landshut, Bavaria, Germany*

- Advisors: Prof. Sandra Eisenreich and Prof. Eduard Kromer
- Topic: 3D object detection with PointPillars algorithm on standard and synthetic LiDAR point cloud datasets

Teaching & Mentoring Experience

Curious Cardinals, *Passion Project & Executive Functioning*, High School Mentor Nov 2025 — Present

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 labs covering Python programming, OOP, debugging, security, and computing fundamentals
- Extracted 20 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 in formal logic, probability and statistics, algorithmic thinking, and simulation, through weekly office hours and personalized feedback on 25 quizzes
- Assisted professors in grading three math and programming assignments per semester

Leadership & Open-source

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

Humane Bench **Evals Contributor** Oct 2025

Selected Research Projects

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

Replication: Counterfactual Test for Faithfulness (github.com/mychiffonn/faithfulness-test) Dec 2025

- Replicated Atanasova et.al. (2023)'s counterfactual test, which can be applied to evaluate whether language model explanations are faithful by checking if perturbing input features leads to changes in model output and explanations (such as chain-of-thought reasoning).

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.

Mnemonic Generation for Vocabulary Learning (github.com/mychiffonn/mnemonic-gen) Oct 2024 — Mar 2025

- Designed an AI chatbot that generated diverse and memorable mnemonic devices for learning and retaining vocabulary, synthesizing 50+ papers across linguistics, psycholinguistics, language education, and large language models
- Utilized chain-of-thought distillation from a teacher model (DeepSeekR1) to instill linguistic chain-of-though reasoning to a student model (Gemma3-1b) through supervised fine-tuning
- Implemented a Direct Preference Optimization (DPO) pipeline for preference modeling on Gemma3-1b using 500 human and LLM-annotated preference pairs (on memorability, imageability, and retention rates)

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

Selected Web/App Projects

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

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 2025 — 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

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, unsloth, trl, petri, scikit-learn, LangGraph, LlamaIndex
- **Web Development**: Astro, React, FastAPI, Flask, Express.js, PostgreSQL, TailwindCSS, shadcn/ui
- **Tools & DevOps**: Git, Docker, Render, Netlify, LaTeX, Zotero, Typst
- **Languages**: Vietnamese (native), English (fluent/C2), Mandarin Chinese (lower-intermediate/band 4)