Tra My (Chiffon) Nguyen

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

Research Interests

Human–AI systems that reflect and support diverse human patterns: **multilingual and multicultural AI**, **AI alignment** (pluralistic alignment, cooperative AI, scalable oversight, evaluation, control, mechanistic interpretability, representation engineering), **interaction-centric AI** (human-AI collaboration, human-AI interaction), and **AI ethics**.

Education

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

B.S 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
- Certificates: Natural Language Specialization (deeplearning.ai, 2023), Machine Learning Specialization (2022)
- Self-study: AI Alignment (ARENA, 2025), Introduction to AI Alignment (Bluedot Impact, 2025)

Research Experience

AI Safety Research Mentee Sep 2025 — Present

Algoverse AI Research (Advisor: Eyon Jang) *Remote*

- Topic: Scaling behavior of chain-of-thought monitoring when threat models sabotage or reward hack

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 pipelie using PointPillars algorithm on standard and synthetic LiDAR point cloud datasets

Causal Inference Research Intern (Replication and Extension) Nov 2023 — Dec 2023

Minerva University (Advisor: Prof. Alexis Diamond) *Remote*

- 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

Teaching & Mentoring Experience

- **Curious Cardinals**, *High School Passion Project*, Mentor Oct 2025 - Present
- **Minerva University**, *PR51 Programming with Python*, Peer Tutor and Data Analyst Spring 2025
- **Minerva University**, *FA51 Algorithmic Thinking and Game Theory*, Lead Teaching Assistant Spring 2023, 2024
- **Minerva University**, *FA50 Formal Logic, Probability and Statistics*, Lead Teaching Assistant Fall 2022, 2023

Open-source

Open-source and Eval Dataset Contributor @ HumaneBench Oct 2025 - Present

- Help build new human-friendly benchmark for AI using Pydantic and Inspect frameworks

Lead Design Engineer @ SEACrowd Aug 2025 - Present

- Main designer, developer and maintainer of seacrowd.org website (Jekyll, Bootstrap)

Dataset Contributor @ Humanity's Last Exam Feb 2025

Selected Projects

Astro 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 (80+) and Lighthouse (100) scores using Astro, TailwindCSS and shadcn/ui

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 other learning retention measures)

Technical Skills

- **Programming Languages:** Python, TypeScript, R, SQL, Bash
- **Machine Learning & Statistics:** PyTorch, trl, unsloth, Inspect, scikit-learn, LangGraph, LlamaIndex, SciPy, PyMC
- **Web:** React, Astro, Express, FastAPI, PostgreSQL, Flask, Jekyll, TailwindCSS, shadcn/ui
- **Tools & Technologies:** Git, Docker, Render, Railway, Netlify, LaTeX, Zotero, Typst