TIANHAO CAO

British Columbia, Canada | snalyf.c@gmail.com | snalyf.github.io | LinkedIn

OBJECTIVE

M.S. Data Science candidate at UBC (Computational Linguistics) with a strong foundation in machine learning, NLP, LLM research, and causal inference. Experienced in building end-to-end ML pipelines, RAG systems, and privacy-focused AI research. Seeking opportunities at the intersection of applied machine learning and language understanding, where rigorous statistical thinking and hands-on engineering can drive real-world impact.

EDUCATION

University of British Columbia British Columbia, Canada
M.S. in Data Science (Computational Linguistics) *September 2025 – June 2026*
Relevant Coursework: Algorithms & Data Structures, Descriptive Statistics & Probability, Data Visualization, Statistical Inference & Computation, Databases & Data Retrieval, Regression, Feature & Model Selection, Parsing for Computational Linguistics, Computational Semantics, Advanced Corpus Linguistics, Computational Morphology, Machine Translation, Natural Language Processing, Deep Learning, Supervised Machine Learning, Unsupervised Machine Learning, Time Series Analysis

University of California, Santa Cruz California, USA
B.A. in Business Management Economics *September 2019 – June 2022*
Relevant Coursework: Advanced Quantitative Methods, Time Series, Machine Learning, Statistics, Econometrics, Corporate Finance, Accounting, Marketing, Security Market, Managerial Cost Accounting

TECHNICAL SKILLS

Programming	Python, R, SQL, MongoDB, Bash
ML & AI	PyTorch, Scikit-learn, XGBoost, Hugging Face Transformers, LLM Fine-tuning, RAG Pipelines, Vector Embeddings, Prompt Engineering, SHAP, VLM
Data Engineering	ETL Pipelines, Feature Engineering, Pandas, NumPy, Data Visualization
MLOps & Cloud	Docker, AWS, Git/GitHub, CI/CD, Weights & Biases (W&B)
3D Modeling	Blender
Languages	English (Native), Mandarin (Native)
Work Auth.	Canadian Post-Graduation Work Permit (PGWP) Eligible

PROJECTS

Membership Inference Attack on Fine-tuned Language Models UBC
Academic Project — Kaggle Competition *GitHub Repository*

- Achieved **AUC 0.907 and TPR@FPR=0.1 of 0.697** on the validation split — a **+55% improvement in AUC** and **+349% improvement in TPR** over the raw-loss baseline — qualifying as the top-performing method in the competition.
- Designed and implemented an Improved Neighborhood Attack using dual-model feature fusion across four models (two fine-tuned SmolLM2 variants at 135M and 360M parameters, plus their base checkpoints), capturing nuanced memorization signals invisible to single-model baselines.
- Engineered a 16-dimensional feature vector per sample spanning five categories: fine-tuned model losses, base model losses, reference ratios, neighborhood perturbation gaps, and cross-model agreement signals — enabling a supervised Gradient Boosting classifier (300 estimators) to learn complex memorization patterns.
- Replaced slow BERT mask-and-fill neighbor generation with a fast token-drop perturbation strategy (10% word dropout, 5 neighbors per sample), dramatically reducing inference time while maintaining perturbation quality; feature importance analysis revealed cross-model reference signals accounted for 86% of model importance.

Membership Inference Attacks on Vision-Language Models Independent
Personal Research Project *GitHub Repository*

- Achieved best result of **AUC 0.849 and TPR@FPR=0.1 of 0.388** by combining loss-ratio features and neighborhood perturbation features via Logistic Regression, progressing from a raw loss baseline of AUC 0.50 across five systematically designed attack variants.
- Adapted the M^{4}I framework (Hu et al., NeurIPS 2022) to a modern CLIP + SmolVLM architecture: replaced legacy ResNet-152/LSTM shadow models with CLIP ViT-B/32 as an off-the-shelf cross-modal feature extractor, achieving Kaggle AUC **0.8056** with no shadow model training required.
- Extended MIA to the multimodal neighborhood setting by generating text perturbations (10% word dropout) and image perturbations (Gaussian noise, patch masking, random crop) across both fine-tuned and base models, extracting 17 cross-modal sensitivity features.
- Identified that CLIP image-text cosine similarity dominates VLM membership signal (48% of XGBoost feature importance), and that VLM-generated text adds near-zero discriminative value (Δ Kaggle AUC = +0.00023 vs. CLIP-only), shifting the attack paradigm from model-behavior to data-alignment.

Full-Stack Corpus Search Engine with LLM Annotation UBC
Academic Project *In Progress*

- Fine-tuned a domain-specific NLP model on Hugging Face for automated corpus annotation, **reducing manual labeling effort by 80%+** while maintaining annotation quality across large text corpora.
- Designed and developed a FastAPI backend with Whoosh-based retrieval indexing, enabling sub-second keyword queries; containerized the full application with Docker and deployed to AWS for reproducible, scalable cloud-hosted access.
- Automated dataset acquisition via the Kaggle API and built structured ETL workflows to prepare raw text data for indexing and annotation; implemented RESTful API endpoints with structured error handling and input validation following production-grade development practices.

Betty Talker — Recipe RAG Pipeline UBC
Academic Project *GitHub Repository*

- Built a full RAG pipeline from scratch: parsed a public domain recipe book (Project Gutenberg) into structured JSON using regex-based text extraction, then generated sentence embeddings with `all-MiniLM-L6-v2` to construct a searchable vector store with cosine-similarity retrieval.
- Designed a few-shot LLM intent classifier to route queries into three actions — recipe search (RAG), ingredient scaling, or off-topic refusal — enabling a controlled and action-aware conversational agent grounded strictly in retrieved context.
- Built a recipe scaling module supporting Unicode fraction characters, computing scale factors from either a direct multiplier or target serving size; packaged the entire pipeline as a modular Python project with a CLI interface and one-command setup script.

Credit Card Default Analysis UBC
Academic Project *GitHub Repository*

- Trained and evaluated multiple classifiers (Logistic Regression, Random Forest, XGBoost) on a 30,000-sample UCI dataset, selecting the optimized XGBoost model which achieved the best Precision/Recall balance (**F1-score: 0.55**) on a heavily imbalanced dataset.
- Conducted extensive EDA to uncover data distributions and class imbalance; applied resampling techniques and feature scaling to improve model robustness across the full end-to-end ML pipeline.
- Utilized SHAP (SHapley Additive exPlanations) values to interpret black-box model decisions, identifying recent payment history as the most critical predictor of default — translating model outputs into actionable business insights.

Predictive Modeling of Agricultural Insurance Purchase Behavior UC Santa Cruz
Academic Project *GitHub Repository*

· Achieved **76% predictive precision** by developing and comparing Lasso Regression (for feature selection) and Random Forest models on a cleaned and merged survey dataset of 1,000+ records with 30+ variables from the Harvard Dataverse.
· Engineered a robust dataset by cleaning and merging raw survey data, then interpreted model outputs via feature importance metrics to translate technical findings into actionable business insights.
· Conducted error analysis to assess prediction variance, discussing model limitations and potential biases in the survey data collection methodology.

Difference-in-Differences Analysis of COVID-19 Policy Impacts UC Santa Cruz
Academic Project *GitHub Repository*

· Implemented a Difference-in-Differences (DiD) causal inference framework in R to quantify the economic impact of early pandemic lockdown policies on GDP across 5 major economies (US, China, Japan, UK, India).
· Aggregated and standardized cross-country economic panel data from the WHO and Our World in Data to construct a time-series dataset; mitigated omitted-variable bias by incorporating control variables to improve robustness and validity of causal estimates.
· Performed robustness checks by comparing the adjusted model against baselines to validate policy implications and discuss potential confounding factors in a real-world policy evaluation setting.