

Eric Wang

Toronto, ON, Canada | yifanwang02@outlook.com | [LinkedIn](#) | [GitHub](#) | [Website](#)

EDUCATION

University of Toronto (U of T), Toronto, Canada Sep. 2022 – Jun. 2027 (Expected)

- **Honours Bachelor of Science:** Data Science Specialist, Statistics Major, and Computer Science Minor
- **Relevant Courses:** Machine Learning, Deep Learning, Data Science, Databases, Time Series Analysis, Probability & Statistics, Data Structures & Analysis, Software Design, Methods of Data Analysis, Survey Sampling, Advanced Calculus, Software Tools & Systems Programming

SKILLS

Core Skills: Data Science, Data Analysis, Data Cleansing, Data Pipelines, Data Engineering, Statistical Modelling, Machine Learning, Model Evaluation, Natural Language Processing (NLP), LLM Applications, Information Retrieval, Data Visualization, Software Engineering

Software and Programming: Python, SQL, R, C, Java, Git, Quarto, MS Office Suite

Data Science: pandas, NumPy, SciPy, scikit-learn, XGBoost, TensorFlow, statsmodels, pygam, PyTorch, Hugging Face Transformers, Sentence Transformers, Model2Vec, Faiss, hnswlib, NLTK, LangChain

Visualization: Plotly, Folium, plotnine, Tableau, Matplotlib

AI Engineering Principles: Prompt Engineering, Context Engineering, Harness Engineering

RESEARCH EXPERIENCE

Correlation-Augmented Neural Retrieval, U of T, Toronto, Canada Sep. 2026 – Present

- Research project, supervised by Prof. Nick Koudas (Department of Computer Science), testing whether word-association statistics (a term-association graph from CRAFT, Corpus Relatedness Analysis Using Fourier Transforms) can help AI search (dense retrieval) cut results that miss a query's topic and find passages that answer it without sharing its words; so far built the baselines and search indexes the experiments will use
- Evaluated and optimized **information retrieval** configurations, measuring embedding throughput, index size, CRAFT pipeline time and memory, and fine-tuning (LoRA) feasibility, enabling efficient experimentation on a laptop
- Matched the model cards' search-quality scores (nDCG@10) within ± 0.01 on all 4 baseline **information retrieval** runs (text embeddings, exact search), generating BGE embeddings for 4 corpora and E5 embeddings for 2
- Met the **model evaluation** target of 0.95 recall@100 on all 6 development indexes, building 8 vector-search indexes (HNSW in **Faiss** and **hnswlib**) and a search implementation whose result-matching checks returned the same top-100 results as Faiss on every development query

WORK EXPERIENCE

Data Operations Engineer (Co-op), AISHU Technology Corp., Shanghai, China May – Aug. 2025

- Developed **LLM applications** (Text-to-SQL, Chat BI) that enabled users to question 3 customer databases in natural language, with **data engineering** (writing and optimizing **SQL**); answers took about 1–13 minutes for questions that could not practically be answered by hand
- Designed about 400 **SQL** test queries for the **model evaluation** (execution accuracy) of the first 2 proofs of concept (POCs), running 9 rounds of business-rule tuning and **prompt engineering** (DeepSeek V3) across all 3; those 2 POCs each improved from about 50% to about 70% execution accuracy
- Turned customer business rules into 30+ Markdown documents (about 6,000+ rows after chunking, across 3 POCs) for the LLM's **context engineering**, with stakeholder management across 5 customer consultations and follow-up, and personally demonstrating to customers the 2 POCs that were delivered

Data Analyst & Logistics Assistant, Marubeni-Itochu Steel Canada Inc., Toronto, Canada May–Aug. 2024

- Produced an **Excel** report for a trend analysis of 5 years of steel imports into Canada, with **data cleansing** and **data analysis** of Statistics Canada data comparing net weights before, during, and after COVID across 10 product categories and 5+ partner countries, and identifying the shifts the analysis used
- Created **data visualization** (charts) and executive summaries for internal stakeholders as part of stakeholder management, maintaining regulatory-change datasets that supported planning and reporting

PROJECTS

My Second Brain by YFW (Python, TypeScript)

Jun. 2026 – Present

- AI knowledge app that runs locally and turns text, documents, images and scanned PDFs (vision transcription), and web pages into a wiki for Obsidian that links each page to its sources, so answers can be traced, and reconciles changed or deleted Markdown incrementally; v1.0.0 passed all 7 citation-grounding checks
- Delivered hybrid **information retrieval** for answers that cite their sources (RAG), fusing keyword (BM25) and vector (LanceDB) results by Reciprocal Rank Fusion within each knowledge scope, adding LLM reranking as an option, and falling back to keyword search when vector search fails
- Deployed v1.0.0 as a **Python** package, command-line tool, and web app (FastAPI, TypeScript, Vite, Pydantic) with OpenAI and Anthropic adapters, passing 176 Python tests in its **software engineering** process

Global CO2 Per-Capita Emissions Analysis with World Bank API (Python)

Jan. – Apr. 2026

- **Data science** project identifying which GDP, energy, forest, and urbanization indicators best predict per-capita CO2 emissions in 196 countries (1990–2023); **XGBoost** gain-based importance ranked log GDP, renewable-energy share, and urbanization as the three highest-importance features
- Built a 7,378-row country-year panel in a **pandas data pipeline**, with **data engineering** (6 World Bank API indicators combined by ISO3 joins), **data cleansing** (region filtering), and feature engineering (log transforms, year centring), keeping 5,499 complete country-year cases
- Reached a test R^2 of 0.722, with error metrics (RMSE 0.543, MAE 0.359), from an **XGBoost machine learning** model under leakage-resistant **model evaluation** (70/15/15 country split, country-grouped 5-fold tuning), comparing it with regression models (OLS, GAM) and a pruned decision tree (R^2 0.706)
- Published the results' **data visualization** as a reproducible **Quarto**/GitHub Pages site with 3 **Plotly** figures, including map animations, distributions by decade, and GDP-emissions views

LLM Source Classification with Survey and Text Features (Python)

Sep. – Nov. 2025

- **Machine learning** project identifying whether ChatGPT, Claude, or Gemini wrote a response, from 825 survey and free-text observations; a feed-forward neural network led on the held-out test set with 69% accuracy (0.69 macro F1)
- Inspected the survey records and, with **data cleansing**, standardized labels and encoded Likert and multi-select answers
- Constructed the feature matrix, combining the survey features with the project's text features (TF-IDF, vocabulary cap) in feature engineering and preprocessing train, validation, and test splits the same way

NLP Classification: Tweet Sentiment and Text Emotion (Python)

Jan. – Apr. 2025

- Text classification of COVID-19 tweet sentiment (3 classes) and the emotion of 18,000 text samples (6 classes); the emotion model chosen in tuning reached 84% test accuracy
- Compared Naive Bayes and softmax regression with **NLTK** feature engineering (word counts, TF-IDF) for **NLP**
- Selected a 2,000-component emotion model (word counts, dimension reduction by PCA, softmax regression) after **model evaluation** with 5-fold stratified cross-validation (83.5% mean validation accuracy), tuning the component count

Global Life Expectancy Regression Analysis (R)

Sep. – Dec. 2024

- Studied how health, economy, education, and development indicators relate to life expectancy (WHO dataset, 1,649 observations), identifying schooling, GDP, adult mortality, and income composition of resources as predictors
- Refined the project's **statistical modelling** in **R**: multiple linear regression with variable transformations, multicollinearity checks (VIF), partial F-tests, and model selection (AIC)
- Validated model performance and made results interpretable, using regression diagnostics and **data visualization** (ggplot2) to ensure statistical assumptions were met

Meal Master (Java)

Sep. – Dec. 2024

- **Java** app that allows users to search, save, and organize recipes, manage accounts, and generate shopping lists (TheMealDB and File.io APIs), with a modular architecture for maintainability and testability; proposed its concept
- Implemented the core of the user and recipe data-access objects with modular **software engineering** design, taking responsibility for the Shopping List use case
- Reviewed code in its **software engineering** process for correctness, tests, architecture, style, and documentation

CERTIFICATIONS

Supervised Machine Learning: Regression and Classification

DeepLearning.AI, Stanford University

Generative AI for Everyone

DeepLearning.AI

LangChain for LLM Application Development

DeepLearning.AI

Building Systems with the ChatGPT API

DeepLearning.AI