

EDUCATION

UNIVERSITY OF MICHIGAN, ANN ARBOR

Master of Science in Computer Science & Engineering

- GPA NA/4.0

Ann Arbor, MI
Aug 2025 – Current

UNIVERSITY OF MICHIGAN, ANN ARBOR

Bachelor of Science in Computer Science & Data Science

- GPA 3.959/4.0 with high distinction
- James B. Angell Scholar
- Core Course: EECS 442, 445, 487, 492; STATS 413, 451

Ann Arbor, MI
Aug 2023 – May 2025

UNIVERSITY OF CALIFORNIA, SANTA BARBARA

Bachelor of Science in Statistics and Data Science

- GPA 3.64/4.0

Santa Barbara, CA
Sep 2021 – Jun 2023

EXPERIENCE

Modeling the Structure and Flow of Arguments in Podcast Discussions Researcher

Ann Arbor, MI
May 2025 – Current

- Under the instruction of Prof. David Jurgens, built an LLM-based pipeline to analyze podcast transcripts, extracting both explicit discussion points and implicit assumptions to study conversational dynamics.
- Developed methods for host-level and cross-speaker analysis, including clustering and visualization of argument flows across episodes.
- Applied embedding models and dimensionality reduction (e.g., t-SNE) to map reasoning styles and reveal implicit patterns in discussions.
- Produced structured datasets and visual outputs for qualitative evaluation.

Faster R-CNN Object Detection Project Group Leader

Ann Arbor, MI
Aug 2024 – Dec 2024

- Developed a custom *VOCDataset* in PyTorch to parse PASCAL VOC2007 annotations into bounding-box tensors and single-class labels.
- Adapted a pretrained ResNet-18 (ImageNet weights) by stripping off its final classification layers to expose a 512-channel feature map suitable for FPN-style object proposals.
- Placed dense anchors at multiple scales and aspect ratios on the feature map, specifically, sizes (32, 64, 128) pixels with aspect ratios (0.5, 1.0, 2.0). This produced a fixed grid of candidate boxes per spatial location, enabling the RPN to regress and score proposals across small, medium, and large objects.

- Bi-linearly pooled variable-sized proposals into uniform 7×7 feature tensors. By automatically selecting the appropriate scale from the backbone's feature pyramid and sampling four points per bin, this module preserved spatial details critical for accurate box refinement and classification.
- Built evaluation routines computing IoU-based detection accuracy and AUROC on both training and validation splits; visualized performance trends across epochs with Matplotlib. Produced high-confidence object detections on unseen VOC images, overlaying bounding boxes on sample outputs for qualitative validation.

QA & RAG project

Ann Arbor, MI

Coder

Jan 2025 – Apr 2025

- Designed and implemented a novel hybrid retrieval strategy that fuses dense vector embeddings with lexical matching and applies Maximal Marginal Relevance (MMR) re-ranking, yielding a 50% relative increase in Exact Match and 47% in F1 over a cosine-only baseline on HotpotQA.
- Extended the EfficientRAG query-optimization pipeline, customizing token-labeling thresholds and filter logic to decompose complex multi-hop questions into targeted sub-queries without repeated LLM calls.
- Built a robust dual-model embedding system, prioritizing a high-capacity Ollama Qwen2.5 model with automatic fallback to MiniLM to ensure consistent performance across computational environments.
- Benchmarked and analyzed three retrieval methods (cosine similarity, MMR, hybrid) in a unified framework, conducting extensive quantitative (EM/F1 metrics) and qualitative error analysis to identify strengths, limitations, and future research directions.

Full-Stack Retrieval-Augmented Chatbot Web App

Ann Arbor, MI

Coder

Jan 2025 – Apr 2025

- Built a Flask server exposing a home route and a /get endpoint that handles chat queries; implemented secure file-upload routes to ingest user-provided PDFs/Markdown into a persistent knowledge_base.
- Designed The_RAG_Process to split and embed documents, storing vectors in Chroma; added a TF-IDF fallback for lightweight similarity search. Created a custom embedding model leveraging the GeminiAI platform, optimizing document similarity searches, and enhancing the chatbot's performance in RAG processes.
- Developed a responsive single-page application with Vue.js and Axios for real-time messaging, using marked.js to render Markdown, JavaScript/CSS for chat bubbles, avatars, and timestamps.

China Merchants Securities International Co., Ltd.

Shenzhen, China

Summer Intern

Jul 2023 – Aug 2023

- Contributed to a team project at this pioneering securities firm, focusing on applying LLMs for public opinion analysis and document analysis, processing over 500 legal documents.
- Rapidly learned and applied new Python packages, leading to a 14.29% increase in project efficiency and enhancing team capabilities in scalable application development.
- Authored and optimized 1500+ lines of Python code for LLM interfacing, which increased data processing efficiency by 7%.