Tony Ding

xiayiding.tony@outlook.com · (213)245-5570 · LinkedIn · Personal Website

EDUCATION

Harvard University, Boston, MA *Aug 2022 – May 2024*
Master of Science Degree in Data Science
- GPA: 3.96/4.0; Relevant coursework: *Deep Learning, Statistical Learning*, *Computing for Big Data*
- Cross-registrations at **MIT**: *Advanced NLP, Advances in Computer Vision, Machine Learning in Healthcare*; GPA: 5.0/5.0

University of Southern California, Los Angeles, CA *Aug 2018 – May 2022*
Bachelor's Degrees in Data Science and in Neuroscience
- Data Science major GPA: 4.0/4.0; Cumulative GPA: 3.9/4.0
- Presidential Scholar (half-tuition awarded throughout 4 years); Renaissance Scholar Distinction

PROFESSIONAL EXPERIENCES

Machine Learning Data Scientist, **Apple**, Sunnyvale, CA *Jan 2026 – Present*
- Developed and fine-tuned multiple **agentic evaluators** for benchmarking Apple Intelligence Planner Agent's performance by using LLMs, assessing agent's **goal completion**, **trajectory optimality**, and **rubrics alignment**. Triaged evaluation results and successfully drove downstream Planner Agent's **system prompt optimizations** and informed **architectural design decisions**.
- Built and fine-tuned a **data evaluation pipeline** that identifies data quality issues and curates high-fidelity multimodal scenario files using **VLMs**, where each scenario simulates a **multi-turn human-agent interaction** of Apple's Visual Intelligence System. Ensures each scenario contains coherent agent goal and trajectory with realistic user utterance and behavior.
- Designed a **user utterance augmentation pipeline** to stress-test Planner Agent robustness across diverse linguistic styles.

Data Scientist, **CVS Health**, Boston, MA *Apr 2024 – Dec 2025*
- Developed and deployed automated case categorization pipelines and a resolution generation **RAG** pipeline using **GPT** family of models, **OpenAI embeddings, LangChain**, and **Airflow DAGs**. Optimized performance via prompt engineering and mitigated LLM hallucination with **self-consistency decoding**. Added **real-time LLM output monitoring** mechanisms using **Ragas** metrics. Led both projects from model ideation and development to production deployment.
- Queried large-scale databases on **GCP** using **BigQuery** and engineered **ETL** pipelines using **Spark** for efficient data retrieval and integration with the automated GenAI pipelines. Built intuitive **Streamlit** Apps for non-technical end users to easily interact with our GenAI functionalities. Further designed and analyzed **A/B tests** to optimize App's UI layouts.
- Benchmarked various LLMs' performances and conducted token cost and model latency analyses to optimize resource allocation and pipeline efficiency. Prepared and presented in-depth project impact analyses for the GenAI pipelines to senior business stakeholders, revealing **$1.7M in total annual savings** and a **532% boost in operational efficiency**.

AI & Data Scientist, **Mayo Clinic**, Boston, MA *Sep 2023 – Dec 2023*
- Deployed an **LLM (PaLM)** on Google Cloud Platform **(Vertex AI)** using **Python** for automating the data abstraction pipeline for the patient management and pathology database.
- Utilized Google BigQuery for data integration and processing. Leveraged Python Regex library and a **semi-supervised topic modeling algorithm**, **Guided Latent Dirichlet Allocation**, to assist LLM in locating the correct text to attend to.
- Conducted extensive clinical **prompt engineering** and LLM **output refinement** strategies to reduce LLM hallucination and to improve overall LLM performance and consistency. Ultimately increased the accuracy of LLM's output by 47%.

RESEARCH PROJECTS

End-to-End Cross-Lingual Summarization (CLS) with Pre-training
Main Affiliation: **MIT Department of EECS;** Topic: **NLP** *Sep 2023 – Dec 2023*
- Developed an **end-to-end framework** for CLS tasks by leveraging **mBART**, eliminating the traditional pipeline approach.
- Achieved a significantly better ROUGE-1 score of 3.82 on an external dataset by **fine-tuning the linear and cross-attention layers** of mBART on CLS data, outperforming the baseline model by 5357%.

Pleural Effusion Diagnosis: Multimodal Approaches Using Deep Neural Nets and Transformer-based Architectures
Main Affiliation: **MIT Department of EECS;** Topic: **NLP + Computer Vision** *Dec 2022 – May 2023*
- Combined patients' diagnosis reports with X-ray images to explore the optimal **fusion strategies** for a **multimodal** approach in diagnosing pleural effusion.
- Researched on early fusion, joint fusion, and late fusion strategies. Fine-tuned VGG16 and DistilBERT using HuggingFace's **PEFT** and **LoRA** configurations and achieved an AUC of 0.9887. Integrated the late fusion multimodal model with an elastic net regularized logistic regression classifier to further enhance classification performance.

RELEVANT SKILLS & ADDITIONAL AWARDS

- Expert in Python and SQL; 3+ years of experience building impactful and efficient Machine Learning and LLM solutions
- Extensive experience in A/B Testing, Natural Language Processing (including RAG pipeline and LLM deployment), Computer Vision, and deep learning algorithms & libraries (LangChain, PyTorch, Tensorflow, NLTK, spaCy)
- Bright Futures Award in the 2023 NNLM Data Visualization Challenge - Complex Visualization Category