YIMING LI

+1 (617) 412-0879 | michaellymtt@gmail.com | Boston, MA

EDUCATION

Boston University 09/2025 - Present
M.S. in Data Science | GPA: 4.0/4.0 Boston, MA

Beijing University of Technology 09/2021 - 06/2025
B.S. in Statistics | GPA: 3.5/4.0 Beijing, China

SKILLS

Programming: Python, R, SQL, C/C++, JavaScript

ML/AI: PyTorch, Scikit-learn, LightGBM, Hugging Face (Transformers, Diffusers), SBERT, Faiss

Data & Tools: Pandas, NumPy, MySQL, MongoDB, Git, Linux, Docker

LLM & Agents: Prompt Engineering, RAG, Multi-Agent Systems (AutoGen, CrewAI, LangChain), LangSmith

EXPERIENCE

Institute of Psychology, Chinese Academy of Sciences, Beijing, China 06/2024 - 08/2025
Machine Learning Engineer Intern

- Architected a high-performance ETL pipeline using Pandas and Python multiprocessing for 100+ GB wearable sensor data, reducing batch processing time from 8 hours to 1 hour (87.5% improvement) across MySQL and MongoDB databases.
- Engineered a hybrid emotion recognition system combining LightGBM for feature selection with a PyTorch MLP classifier, achieving 5% F1-score improvement on 50K+ multimodal physiological samples (heart rate, skin conductance, accelerometer).
- Built an interactive psychology experiment platform in Python (Pygame) featuring scenario-based emotion elicitation and automated subjective report collection, enabling 3x faster experiment iteration and parallel participant sessions.
- Collaborated with cross-functional team of 5 researchers and 3 developers to build real-time stress intervention mobile app; designed and validated stress detection algorithm achieving 85% precision in controlled user trials.

Jiahua Information Technology, Beijing, China 01/2024 - 03/2024
Software Development Intern, Product R&D Department

- Built end-to-end NLP pipeline for banking conversational AI: applied prompt engineering to generate 10K+ synthetic dialogues from LLM, then trained and deployed BERT-based models for intent classification (92% accuracy) and dialogue summarization.
- Performed data cleaning and annotation on raw banking dialogue corpus; applied clustering analysis to identify conversation patterns and customer intent categories, improving training data quality for downstream NLP models.

PROJECTS

Colombian Extradition Data Agent (Client Project) | Python, LangChain, AutoGen, LangSmith 09/2025 - 12/2025

- Architected multi-agent information extraction system for 3.3K+ legal documents; designed single-agent prototype achieving 94% extraction accuracy (vs. 72% baseline regex) before scaling to production multi-agent architecture.
- Benchmarked orchestration frameworks (AutoGen, CrewAI, LangChain) for tool-call reliability; integrated LangSmith for agent trajectory tracing, improving inter-agent communication success rate by 25%.

Deep Learning Applications | PyTorch, Transformers, Diffusion Models, Hugging Face 10/2025 - 12/2025

- Built decoder-only Transformer for symbolic math: implemented custom tokenizer and attention mechanism from scratch to evaluate arithmetic expressions with nested parentheses, achieving accurate step-by-step computation on 5+ operand expressions.
- Developed DDPM-based image generation system on 22K animal images (11 classes); optimized U-Net architecture and noise scheduling to produce diverse 64×64 outputs, achieving Inception Score of 8.4 across all categories.

Epidemic Intelligence Data Pipeline | Python, Docker, OpenShift, PostgreSQL, REST APIs 01/2026 - Present

- Building data ingestion pipeline for health surveillance system: developed containerized API collection services with Docker, implemented data validation and cleaning workflows, and deployed on OpenShift cluster.

Neural Decision-Making Modeling (Undergraduate Thesis) | Python, MATLAB, Bayesian 10/2024 - 06/2025

- Applied computational neuroscience methods, using point process models to analyze neural spiking activity from the orbitofrontal cortex to quantify and decode neural signals during economic decision-making.
- Developed a novel state-space model for the interleaved learning paradigm to address the challenge of analyzing complex learning behaviors. This model simultaneously integrates continuous data and discrete data, using Bayesian methods to dynamically track latent cognitive states, providing a new quantitative analysis framework for the field.