Aya Oshima

ayaoshima.us@gmail.com • linkedin.com/in/ayaoshima • github.com/aya0221 • ayaoshima.me • New York, NY (open to relocate)

SUMMARY

Machine Learning Engineer (M.S. CS, NYU) with hands-on industry experience deploying real-time AI systems and a research background in computational neuroscience. Specializes in NLP (speech recognition, NER, LLMs) and deep learning pipelines, delivering high-accuracy, low-latency solutions for real-world applications. Proven ability to fine-tune transformer models and build end-to-end ML systems (from embedded voice-driven robotics to scalable voice-NLP recommender systems with low latency).

SKILLS

- **Programming & Frameworks**: Python, PyTorch, TensorFlow, Hugging Face Transformers, spaCy, scikit-learn
- **MLOps & Systems**: MLflow, Docker, Linux, Git, CI/CD, AWS, OpenSearch, Agile methodologies
- **Expertise**: Natural Language Processing (NER, ASR, LLM), Computer Vision (CNNs), Deep Learning, Real-Time Inference, Semantic Search, Statistical Modeling (PCA, TCA), Data Visualization (Matplotlib)

PROJECTS

Low-Latency Voice-to-NLP Recommender – Real-time Voice AI System (Apr 2025 – Present)

- **End-to-End Voice Interface:** Engineered a real-time voice-driven recommendation system delivering personalized content from spoken queries via a FastAPI backend and React frontend.
- **Transformer-Based NLP**: Integrated Whisper ASR for speech-to-text and fine-tuned transformer models (DistilBERT for intent classification, RoBERTa-based spaCy NER for entity extraction) on local GPU, achieving 100% intent classification accuracy and 99.97% entity-recognition F1 on noisy speech input.
- **Semantic Search & Optimization:** Implemented semantic search with OpenSearch to retrieve relevant results; optimized the pipeline with batched GPU inference and asynchronous processing for <0.2s end-to-end response latency. Tracked experiments and model versions using MLflow to ensure reproducibility.

Tensor Component Analysis for Neural Plasticity – M.S. Thesis Research (Sep 2023 – Aug 2024)

- **Deep Learning Pipeline:** Developed a Tensor Component Analysis (TCA) workflow in Python and MATLAB to decompose high-dimensional neural spike-train data into core components, enabling significant reduction in dimensional complexity and extraction of structured firing patterns from neuron activity.
- **Neuroscience Insights:** Analyzed oxytocin-driven learning in hypothalamic neuron data (maternal behavior context), uncovering interpretable latent dynamics. Identified how synchronized neuron firing patterns emerge with experience-driven plasticity, providing insights into neural network behavior during learning.
- Conducted research at NYU Grossman School of Medicine, Neuroscience Institute, under Prof. Robert Froemke.

Neural-Symbolic VQA – Multi-Modal Visual Reasoning System (Sep 2022 – Dec 2022)

- Built a multi-modal Visual Question Answering system by integrating CNN-based visual processing, GRU-based language understanding, and symbolic rule reasoning; trained on the Sort-of-CLEVR dataset (10k images, 200k QA pairs), achieving 88% accuracy on relational and 99% on non-relational queries.

EXPERIENCES

NYU Center for Neural Science – *AI Researcher* (New York, NY | Jun 2023 – Aug 2023)

- Developed DeepDream-based generative models using ResNet & VGG16 CNNs to synthesize complex visual stimuli, automating image creation for primate cognitive experiments in reinforcement learning studies.

Nihon Business Data Processing Center – *AI Robotics Engineer* (Kobe, Japan | Feb 2021 – Jul 2022)

- Built and deployed an embedded voice-to-motion AI system, showcased live to 10K+ exhibition visitors. Integrated offline speech recognition (Julius, SRILM) and text-to-speech (OpenJTalk) with NLP parsing and real-time motion control, enabling seamless voice-commanded robot demonstrations on stage.
- Built NLP models for a future voice chatbot product by fine-tuning large pre-trained models (BERT, GPT-2, BlenderBot) and designing custom lightweight LSTM architectures.

EDUCATION

New York University, Tandon School of Engineering (New York, NY) – M.S. in Computer Science, 2022–2024