

# WONHA SHIN

+1) 551-245-2356 | [leahnote01@gmail.com](mailto:leahnote01@gmail.com) | Portfolio: [leahnote01.github.io/](https://leahnote01.github.io/) | LinkedIn: [linkedin.com/in/wshin7](https://linkedin.com/in/wshin7)

## PROFILE AI Engineer | Applied ML & NLP & LLM | Cloud MLOps

Machine Learning Engineer & Data Scientist specializing in **Applied NLP, LLMops, and Cloud MLOps systems**. Experienced in the **end-to-end ML lifecycle** — from large-scale data ingestion and model training to **real-time inference, observability, and automated retraining**.

Proficient in **Python, PyTorch, and cloud-native pipelines (AWS, Azure, Databricks)**, with strong expertise in **LangChain, Kafka, FastAPI, and Hugging Face Transformers** for building scalable, production-grade AI systems. Passionate about **bridging research and deployment**, designing **LLM-powered solutions** that deliver measurable impact in real-world environments.

## SKILLS & PROJECTS

**Programming & Tools:** Python (NumPy, Pandas, FastAPI, Seaborn), R (RStudio), Git, Shell Scripting, REST APIs

**Data Engineering & Streaming:** PySpark, Databricks, Delta Lake, Apache Kafka, Apache Airflow, Spark Streaming

**Cloud Platforms:** AWS (S3, EC2, ECR, Lambda, IAM), Azure (ACR, ACI, Azure ML), GCP (Vertex AI, BigQuery)

**Databases & Data Management:** SQL, PostgreSQL, MySQL, Data Warehousing, ETL Pipelines, Data Integration

**Deep Learning & NLP:** PyTorch, TensorFlow, Keras, XGBoost, Hugging Face, LLMs, LangChain, OpenAI API

**MLOps & Deployment:** MLflow, Weights & Biases (W&B), Docker, CI/CD (GitHub Actions), Prometheus, Grafana

- **LLM-Based Real-Time Translation System with MLOps Automation** | *Kafka • FastAPI • Hugging Face • W&B:*

Built a **low-latency real-time translation system** using Kafka and FastAPI microservices with Hugging Face Transformers in PyTorch, processing 10K+ concurrent requests. Deployed **containerized models** with Docker and automated CI/CD for scalable inference and version control. Integrated **Prometheus/Grafana** for monitoring and **W&B** for experiment tracking, drift detection, and retraining automation.

- **RAG-Powered LLM Application Development** | *LangChain • OpenAI • FAISS • Streamlit*

Engineered an **end-to-end RAG system** using **LangChain** and **OpenAI GPT** for contextual Q&A and **AI Agent-driven task orchestration** over enterprise documents. Built **FAISS-based vector retrieval** and modular **retrieval chains**, deployed via **Streamlit**, and optimized inference with **caching and API parallelization** for scalable, production-grade performance. Enhanced retrieval quality through **embedding evaluation, prompt refinement, and agent coordination**, applying **MLOps-driven design** for reproducible LLM pipelines.

- **Real-Time Tweet Sentiment Analysis Pipeline** | *Databricks • PySpark • AWS • Delta Lake*

Developed a **structured-streaming pipeline** on AWS using PySpark to classify and monitor live tweet sentiment in real time. Designed **Bronze/Silver/Gold Delta architecture** for scalable ETL and optimized partitioning for **low-latency, fault-tolerant processing**. Created KPI dashboards and model performance tracking using Databricks.

- **Microsoft Azure MLOps Pipeline Implementation** | *Azure ML • Docker • GitHub Actions • ACR/ACI*

Developed a **CI/CD automation pipeline** for ML model training and deployment using Docker and GitHub Actions. Integrated Azure Container Registry (ACR) and Azure Container Instances (ACI) for containerized workflows ensuring reproducibility and traceability.

- **E-Cigarette Perception Analysis (Capstone)** | *Python • NLP • scikit-learn • SQL • Tableau*

Led **NLP analysis** on 100K+ **multilingual** social posts to identify misinformation and health-risk patterns. **Fine-tuned BERTweet** and Twitter Twin BERT Large achieving >94% accuracy for relevance and attitude classification. Conducted **topic modeling** with BERTopic and LDA, revealing cross-linguistic differences in health perception, lifestyle, and policy attitudes. Findings are being prepared for publication to inform public-health communication strategies.

- **Cross-Cultural NLP Analysis of European Hotel Reviews** | *Python • LDA • Data Visualization*

Analyzed over 500K multilingual hotel reviews from European countries to study **regional sentiment trends**. Applied **TF-IDF** and **LDA** for topic extraction, revealing cultural variations in customer experience perception across languages.

- **200 Days Learning & Knowledge-Sharing Challenge (2024–2025)**

Completed a **200-day continuous ML & MLOps learning initiative** documenting daily technical studies on GitHub and LinkedIn. Covered topics spanning **statistics, model evaluation, drift detection, deployment, and system monitoring**, strengthening foundations across theory and production-level implementation.

PROFESSIONAL EXPERIENCE

**BSGH Lab — Behavioral, Sexual, and Global Health Research Lab**  
*Research Internship*

**Remote**  
*Jul 2024 – Present*

Collaborating with U.S. public-health researchers to analyze multilingual data on health communication and behavioral patterns. Contributing to research manuscript preparation through data preprocessing, visualization, and clear technical documentation for interdisciplinary teams.

**VOESH New York**  
*Administrative Accounting Specialist*

**South Plainfield, New Jersey**  
*Oct 2018 – Mar 2023*

Built a **predictive sales-forecasting model** in Oracle SQL, improving forecast accuracy by 15% across 10K client accounts. Automated monthly financial dashboards integrating sales and inventory data to support executive decision-making. Partnered cross-functionally with leadership and operations to translate data insights into actionable business strategies, strengthening communication between finance and management.

**EASTERN AMERICAN CDC (Non-Profit Federal Community Bank)**  
*Credit Analyst Intern*

**Englewood, New Jersey**  
*May 2018 – Oct 2018*

Conducted quantitative assessments of 100+ SBA loan applications to identify risk and performance trends. Designed variance and trend analyses to improve portfolio risk evaluation and lending strategies.

EDUCATION

**University of Rochester**  
Master of Data Science GPA 3.43 / 4.0

**Rochester, New York**  
*Aug 2023 – Dec 2024*

- Focus: Machine Learning, Natural Language Processing, Data Engineering, MLOps
- Capstone: Exploring E-Cigarette Perceptions Across Languages Using Transformer-Based Models (Manuscript in Preparation)

**Pusan National University**  
Bachelor of Arts in International Business and Economics GPA 3.73 / 4.0 (**Cum Laude**)

**Busan, South Korea**  
*Mar 2011 – Aug 2018*

- Completed honors-level coursework in econometrics, quantitative analysis, and international finance.