Hudanyun Sheng

hudanyun.sheng@outlook.com | (+86) 13683132915 | https://danniesheng.github.io/ | Beijing, China

CORE SKILLS

- Machine Learning & GenAI: classical ML (classification, ranking, feature engineering), Retrieval-Augmented Generation (RAG), prompt engineering, LLM evaluation, embedding-based retrieval, cross-encoder reranking, uncertainty handling
- Data & System Engineering: end-to-end ML / GenAI pipeline design with quality gating and fallback strategies, noisy document understanding (PDF parsing, OCR, table extraction), human-in-the-loop systems for review, validation, and error correction
- Platform & Cloud: Databricks-based ML workflows and feature pipelines, containerized deployment with Docker, Azure-based ML / LLM platforms in regulated environments, cloud-native data workflows
- Tools & Stack: Python, Pandas, NumPy, scikit-learn, FAISS, Hugging Face (E5 embeddings, rerankers), OpenAI / Azure OpenAI APIs. Kedro, Streamlit

PROFESSIONAL EXPERIENCE

Johnson & Johnson - **Data Scientist (Contractor)** Center of Excellence Beijing, China | April 2024-present

Commercial ML System for Drug Launch & HCP Targeting

- Built early-adopter prediction and HCP ranking models using historical sales, promotion, and engagement data
- Designed feature engineering and scoring pipelines to support large-scale prioritization under business constraints
- Implemented modular Kedro-based data pipelines with validation and reproducibility controls

GenAI / RAG Systems for Medical Content Governance (RAG)

- Designed and prototyped retrieval-augmented generation (RAG) workflows to detect off-label or non-compliant content by comparing localized medical FAQs with regulatory drug labels (Japanese market)
- Led prompt design, retrieval strategy selection, and error analysis to handle ambiguous translations and partial matches between source and target documents
- Combined embedding-based retrieval, rule-based checks, and LLM reasoning with fallback gating mechanisms

Medical Knowledge Retrieval & Doctor Lookup Platform

- Built multilingual medical RAG system integrating FAISS, E5 embeddings, cross-encoder reranking, and uncertainty gating
- Developed robust name normalization and fuzzy matching (NFKC, kana–romaji conversion, CN/JP unification) to improve recall across writing systems
- Built Streamlit UI for evidence tracing, parameter tuning, and reviewer validation

Document Understanding & PDF Table Extraction Pipeline

- Designed resilient pipeline for extracting hospital schedule tables from heterogeneous PDFs (digital, scanned, malformed)
- Integrated Camelot, OCR-based detection, and LLM-based image reasoning with quality gating and fallback logic
- Implemented failure-aware handling for encoding errors, broken PDFs, OCR degradation, and oversized images to prevent pipeline interruption
- Developed reviewer-facing UI for batch validation and annotation

Zenni Optical - **Data Scientist (**AI/M) Beijing, China | June 2023-Jan 2024

Prescription (Rx) Recognition System

- Built FastAPI-based inference service to extract structured optical parameters (sphere, cylinder, axis) from user-uploaded images

- Integrated OCR outputs with rule-based and heuristic post-processing
- Adapted parsing logic for Japanese prescription formats to support multilingual layout variation
- Integrated CloudSQL and BigQuery for monitoring and multi-region usage analytics

Human-in-the-Loop Evaluation & Quality Diagnostics

- Built internal Streamlit tool to visualize OCR and extraction outputs for rapid debugging
- Analyzed API logs and user behavior to diagnose extraction failures related to image quality and capture issues

University of Texas Southwestern Medical Center - **Data Scientist**

Quantitative Biomedical Research Center Dallas TX USA | Sep 2021-May 2023

CyTOF Imaging & Spatial Data Integration System

- Designed end-to-end pipeline integrating multiplexed imaging with single-cell expression data
- Built Flask-based service with Dockerized deployment and parallelized computation
- Achieved ~10× runtime speed-up for large-scale experiments

Digital Pathology: Nuclei Segmentation & Classification

- Developed Mask R-CNN–based pipeline for nuclei segmentation and 6-class classification on H&E slides
- Customized loss functions for partially labeled data, recovering ~20% of training samples
- Achieved 82.5% detection and 82.0% classification performance under weak supervision

Donald Danforth Plant Science Center - **Data Science Researcher**

Data Science Facility St. Louis MO USA | Feb 2020-Sep 2021

Multimodal Imaging & Analysis Systems for Plant Phenotyping

- Built automated pipelines for large-scale multimodal plant imaging (RGB, thermal, hyperspectral)
- Developed instance-level segmentation and temporal tracking workflows for growth analysis
- Contributed to the open-source PlantCV project (modules, tests, documentation, community support)
- Supported interdisciplinary teams with reproducible analysis and visualization

PERSONAL PROJECTS

AI News Agent – LLM-based News Summarization and QA Assistant

Open-source project: daily AI news aggregation, summarization, and question answering

- Designed and implemented an end-to-end LLM-powered news agent for daily AI news aggregation, summarization, and question answering
- Built a lightweight Retrieval-Augmented Generation (RAG) pipeline integrating web scraping, local embedding models, hybrid retrieval, and OpenAI APIs
- Implemented scheduled crawling, automatic topic tagging, and corpus-based QA over historical news data
- Developed interactive Streamlit front-end and deployed on Hugging Face Spaces; explored Docker-based deployment and embedding model switching

EDUCATION

University of Florida - *M.S. in Electrical and Computer Engineering (GPA: 3.86/4), December 2019*

Master thesis: Switchgrass Genotype Classification using Hyperspectral Imagery

University of Florida - *M.S. in Industrial and Systems Engineering (GPA: 3.87/4), December 2017*

Tongji University - *B.S. in Physics (GPA: 4.45/5), June 2015*

Bachelor thesis: The Correction of the Intensity Unevenness of X-Ray KB Imaging