Handi (Andy) Xie

handi.xie.beintouch@gmail.com | 330-247-8892 | linkedin.com/in/handi-xie | github.com/palmpalm7

EDUCATION

M.S. in Data Science   **May. 2023 - Dec. 2024**
Boston University   *Boston, MA*
B.S. in Computer Science   **Aug. 2016 - Dec. 2020**
University of Illinois at Urbana-Champaign   *Champaign, IL*

PROFESSIONAL EXPERIENCE

Data Scientist Intern   **May. 2024 - Nov. 2024**
Office of the CTO, Red Hat   *Boston, MA*
- Designed document pre-processing tool for InstructLab (github.com/instructlab)'s retrieval pipeline. Tuned models achieved over 3% increase on internal TruthfulQA dataset and 7% increase on MMLU.
- Implemented Graph RAG proof-of-concept for InstructLab tuned LLMs' downstream tasks on RedHat OpenShift with Neo4j, **increasing retrieval helpfulness by 9.5% on internal benchmarks.**

Data Scientist (Full-time)   **Mar. 2021 - Dec. 2022**
TanTan, Hello Group Inc.   *Beijing, China*
- Gathered and cleaned user behavior data with HiveQL, performed feature engineering, built ensembled deep learning models in Python to predict churn users and user life-time values, with over 90 % AUC-ROC and F1.
- Designed quasi experiment using casual inferences for online-advertisement scenarios in replace of A/B Test.
- Designed data pipeline and constructed data dashboards. Conducted cluster analysis on anomalies**.**

Project Manager Intern   **Sept. 2023 - May. 2024**
X-lab, BU Spark!   *Boston, MA*
- Concurrently managed 8 data science teams for various stakeholders on challenging civil and social projects.
- Conducted weekly stands-up with Scrum, designed the strategy to drive the product from preview to GA.

PROJECTS

InstructLab (model-agonistic open-sourced tuning framework)   **May 2024 – July 2024**
- Designed and implemented document-specific and semantic-search enabled knowledge source chunking.
- Automated End-2-End model validation pipeline for InstructLab command line interface.

Digitalization of Herbaria (collection of dried-plants)   **Jan. 2024 – May 2024**
- Architected an end-to-end MLOps pipeline that automated data scraping and pre-processing of herbaria data from Chinese Virtual Herbarium, fine-tuned with both local TrOCR model and commerical OCR model on handwritten Chinese herbaria with PyTorch, achieving over 45% accuracy improvement over baseline. SOTA.
- Capsulized and orchestrated scraping, training, validation, automated reporting with docker and Airflow.

Lean Marketing: Info-stream Advertisement Optimization   **Feb. 2022 - Dec. 2022**
- Architected Tantan's lean marketing strategies using RTA (real-time marketing APIs) with deep learning.
- Reduced 5% on CAC (Customer Acquisition Cost) for second-day retention users, over 15% on CAC for the weekly retention users on info-stream advertisements. **Saved marketing budget over $150k/month**.
- Led cross-departmental collaboration with marketing & growth and information security.

Tantan Experiment Platform   **June 2021 - Aug. 2021**
- Conceptualized, architected, and implemented experimental science methods (Propensity Score Matching, Synthetic Control Methods, CUPED) for Tantan's Experiment Platform, reducing reopen rate to 3%.
- Built evaluation methods for content operations and recall advertisements; summarized comprehensive insights that enabled the management team to prevent loss by 12% in six months rapidly.

TECHNICAL SKILLS

- **Languages**: Python (Airflow, Flask, Selenium, PyTorch, LangChain, LLM Sherpa, DSPy), Go, SQL, R, JS
- **Tools**: Git, Docker, Kubernetes, OpenShift, Linux, HuggingFace, Azure Google Cloud, Tableau, Excel
- **Skills**: LLM fine-tuning, AI Agents (RAG, CoT, ReAct, Prompt Engineering and Optimization), Machine Learning (Logistic Regression, Gradient Boosting Trees), Deep Learning (RNNs, CNNs, GNNs, Transformers, Diffusion), CV(Object Recognition and tracking, Segmentations), Big Data (Spark, Hadoop, Flink), A/B Testing (Randomized Experiment Design, Casual Inference quasi-experiment), Data Visualization, Product Analysis.