

# Ashtik Mahapatra

New York • [ashtikkm@gmail.com](mailto:ashtikkm@gmail.com) • +1 (716) 573-8945 • [linkedin.com/in/ashtik-mahapatra](https://www.linkedin.com/in/ashtik-mahapatra)  
[github.com/fieryash](https://github.com/fieryash) • [fieryash.github.io](https://fieryash.github.io)

Experienced Data Scientist skilled in NLP, computer vision, and generative AI, with award-winning ML pipelines deployed at Wolters Kluwer generating millions in revenue. IEEE-published researcher proficient in transformers, diffusion models, end-to-end system development, cloud automation (AWS/Azure), and building production-grade AI for accessibility, legal analytics, and diverse applications.

## EDUCATION

**Masters of Science in Computer Science & Engineering (AI/ML)** Aug 2024 - Jan 2026  
University at Buffalo (SUNY Buffalo) GPA: 3.92

## WORK EXPERIENCE

**Data Science Intern** Jul 2025 - Sep 2025  
**Magnit** United States

- Designed and deployed end-to-end **RAG agent pipelines**, combining vector databases, hybrid semantic-keyword retrieval, and reranking, to automate business-license processing and power document-scale QA, achieving **85% relevancy** and substantial efficiency gains across large unstructured datasets.

**Data Scientist II** Aug 2021 - Jun 2024  
**Wolters Kluwer** Pune, India

- Borrower Analytics – Collateral Intel (Patent-pending, >\$6 M revenue)**: Architected a pipeline that ingests **22 M+ UCC lien filings** (50 states, 23 yrs), classifies **20+ form types with DONUT (98 % F1)** and extracts key fields with **LiLT + regex (90 % F1)**; prototyped open-source LLM workflows for structured data extraction and built monitoring/QA dashboards.
- LegalView BillAnalyzer (>\$20 M revenue)**: Fine-tuned models and rule logic to onboard new clients/law firms, boosting **guideline-matching precision to 80 % and recall to 70 %**; engineered a rate-management system handling **1 M+ records/min** and replaced legacy classifiers with spaCy/FastText, **raising precision by 14 pts and recall by 10 pts** while migrating the module to Dataiku for full MLOps governance.

**Strategy Intern** Feb 2021 - Aug 2021  
**IBM** India

- Built automated pipelines using Step Functions and CloudFormation to tie S3, EMR and Redshift for the backend of an upcoming Asset Management business. Researched AI strategies and integration opportunities in Asset Management.




**Data Science Intern** Oct 2020 - Feb 2021  
**Lawnics** India

- Created Named Entity Relationship Recognizer using Spacy and BERT for Information Retrieval of Legal Documents and achieved an accuracy of **85%+ in 5 target values** (Cases, Sections, Acts, Articles and Citations) and implemented a model for ranking of documents based on similarity index of the query using BM25 and indexed them into ElasticSearch.

## AWARDS

**Winner by Global Innovation Awards** Jan 2023  
Winner of Global Innovation Awards 2023 for the Borrower Analytics - Collateral Intel platform.

## PROJECTS

- Citi Bike Prediction** 
- Built a full-stack MLOps pipeline that ingests Citi Bike trip data, engineers lag features, trains LightGBM models (Hopsworks + MLflow), and serves 24-hour station-level ride forecasts via Streamlit—automated end-to-end with GitHub Actions CI/CD.
- Memory-Augmented Multi-Hop QA System on HotpotQA** 
- Built a **multi-hop QA system on HotpotQA** using a Memory-Augmented Neural Network with BERT embeddings, dual LSTM controllers, and a Neural Turing Machine-style memory for reasoning across paragraphs. Integrated BM25/BERTScore for paragraph retrieval and optimized training with mixed precision and learning rate scheduling. **Achieved 62.5% Exact Match and 90.5% F1 on the validation set.**
- Text-to-ASL Video Generation using Diffusion Models** 
- Engineered an end-to-end deep-learning pipeline that turns text into fluent, sentence-level ASL videos: LoRA-tuned Stable Diffusion for gesture fidelity, Mediapipe skeleton alignment for consistency, and custom frame-blending logic for smooth motion.

## SKILLS

**Programming Languages:** Python  
**Frameworks and Tools:** PyTorch, TensorFlow, NumPy, Scikit-Learn, SpaCy, FastText, Dataiku, Apache Solr, ElasticSearch  
**Machine Learning:** NLP, GenAI, Computer Vision  
**Cloud Technologies:** AWS, Azure, Dataiku, Docker, GitHub

## PUBLICATIONS

**A Novel Approach for Identifying Social Media Posts Indicative of Depression**  on IEEE Jan 2021