PRATHAMESH PRAVIN MORE

prpmore@gmail.com (930)-333-4542 linkedin.com/in/more-prathamesh github.com/Spidey13 Publications

PROFESSIONAL SUMMARY

Recent M.S. Data Science graduate (Indiana University, GPA: 3.6) specializing in Generative AI, Large Language Models (LLMs), and end-to-end MLOps. Proven ability to architect and deploy scalable machine learning solutions, from multi-modal analysis to high-impact predictive modeling. Seeking a challenging AI/ML Engineer or Data Scientist role to build cutting-edge models using Python, PyTorch/TensorFlow, and AWS.

EDUCATION

Master of Science in Data Science Indiana University, Bloomington, IN *05/2025*
Relevant Coursework: Computational Linguistics, LLMs, Big Data Economics, Statistics, Data Mining, Data Visualization
Bachelor of Technology in Computer Engineering University of Mumbai, Mumbai, India *07/2023*
Relevant Coursework: Data Structures and Algorithms, Big Data Analytics, Artificial Intelligence and NLP

PROFESSIONAL EXPERIENCE

Data Science Intern *Indiana University — Bloomington, IN* *05/2024 to 08/2024*

- Architected advanced topic modeling pipelines using **LDA, NMF, K-Means, and Spectral Clustering**to analyze **550+ architectural and design publications** for **DEI themes**achieving **93% Topic Coherence** and developing a modular text-mining framework whose findings were published in *The Design Journal (Taylor & Francis, Feb 2025)*.
- Engineered interactive **Plotly and Dash** dashboards, for temporal analysis of **150+ DEI-related terms**, with dimensionality reduction and **word co-occurrence analysis** via and visualize emerging semantic trends.
- Collaborated extensively with **domain experts** to iteratively refine feature selection strategies and the interpretation of topic modeling outputs, significantly enhancing the granularity and actionable insights derived from design industry literature.

Machine Learning Engineer Intern *Dimensionless Technologies — Mumbai, India* *05/2023 to 07/2023*

- Devised an **NLP pipeline** using **BERT** with AWS, achieving precision of **0.84** to continuously analyze articles from CNN and update a **real-time dashboard** with categorized insights for **trend monitoring and market analysis**.
- Implemented an **OCR model** for industrial documents and deployed a **QA system** that improved **information retrieval speed by 70%**, automating data extraction and streamlining the Tender Document creation process.
- Partnered with clients and cross-functional teams to refine **NLP models** based on business needs and contributed to **data strategy, model interpretability, and deployment discussions**.

Full Stack Developer Intern *Benchmark Computer Solutions Pvt. Ltd. — Mumbai, India* *08/2022 to 10/2022*

- Built a **Flask-based web application** to parse and display structured data from **1K+ uploaded resumes**, exporting results to CSV and reducing candidate screening time by **30%**.
- Implemented **SpaCy, Transformers**, and rule-based techniques to extract fields like skills, education, and experience, improving entity extraction F1-score from 0.78 to 0.91 and cutting annotation cycles by 40% via active learning.
- Developed a **RESTful API** to streamline resume ingestion, supporting scalable backend processing workflows.

SKILLS

Programming Languages & Data Tools: Python, R, SQL, Java, C++, Scala, MySQL, PostgreSQL, MongoDB, Firebase, Pinecone, Spark, Hadoop, Airflow
Machine Learning & NLP: Transformers, LLMs, LSTMs, GRUs, CNNs, RNNs, YOLO, SMOTE, Feature Engineering, Hyperparameter Tuning, PCA, t-SNE, SHAP, Named Entity Recognition, Sentiment Analysis, Topic Modeling, Text Summarization, OCR, SpaCy, NLTK, LangChain
Cloud & MLOps: AWS (S3, SageMaker, ECS, Lambda), Azure (Blob Storage, Cognitive Services), MLflow, DVC, Kubeflow, Docker, Kubernetes, GitHub Actions, CI/CD, FastAPI, REST APIs
Frameworks & Visualization: TensorFlow, PyTorch, Scikit-Learn, NumPy, Pandas, SQLAlchemy, Flask, Django, Matplotlib, Seaborn, Plotly, Grafana, NetworkX, Power BI, Tableau

Projects

Audio-Lyric Emotional Alignment — OpenL3, SentenceTransformers, Librosa, PCA, K-Means, NLTK, Matplotlib

- Built a pipeline using the **MuLan multimodal model** on the **DEAM dataset** (**1.8K+ songs**) to analyze emotional alignment between music and lyrics, applying **PCA** for dimensionality reduction and computing **Multimodal similarity scores** (cosine similarity between normalized audio and text embeddings).
- Analyzed emotion trends across **5 major genres** using **K-Means, Silhouette Scores, and t-SNE**, revealing genre-specific sentiment patterns to improve emotion-aware music recommendation systems.
- Integrated **LangChain** with LLMs to generate concise, context-rich summaries of song lyrics, improving tone interpretation and supporting downstream emotion classification with **92% coverage accuracy**.

F1 Race Strategy Simulator — CatBoost, XGBoost, MLflow, DVC, Docker, FastAPI, MongoDB, Streamlit, Plotly

- Engineered a **high-fidelity Digital Twin** of F1 race dynamics, leveraging **CatBoost and XGBoost** on high-frequency telemetry data to accurately model and predict critical variables like lap times and tire degradation curves.
- Developed a **multi-agent simulation environment** to test and optimize race strategies, exploring reinforcement learning principles to identify optimal pit stop timings and driver tactics under various competitive scenarios.
- Operationalized the full simulation and inference pipeline using **MLflow for versioning** and **Docker for deployment**, delivering actionable strategic insights to a **Streamlit-based Decision Intelligence dashboard**.

Wafer Fault Detection — Flask, MongoDB, Random Forest, XGBoost, OpenCV, PCA, Scikit-learn, MLflow, DVC

- Engineered a high-precision **anomaly detection system** for semiconductor manufacturing, leveraging **XGBoost and Random Forest** on high-dimensional sensor data (200+ features) to classify wafer defects with a **94% F1-score**, directly contributing to production yield optimization.
- Architected and operationalized an **end-to-end MLOps pipeline** using **MLflow, DVC, and Docker**, enabling automated, reproducible model training and deployment cycles that slashed model retraining time by **10x**.
- Integrated a **real-time model monitoring solution** with proactive **data drift detection (97% sensitivity)**, ensuring sustained production accuracy and triggering retraining alerts via a **Flask/MongoDB-backed dashboard**.

Text Summarizer — Python, Flask, HuggingFace, TensorFlow, Scikit Learn

- Authored and implemented a novel extractive summarization algorithm, published in **IEEE CONIT 2023**, combining a **Query-Based Summarizer (QBS)** with **Kneser-Ney smoothing** to generate grammatically sound and contextually relevant summaries.
- Benchmarked the algorithmic approach against state-of-the-art abstractive models by fine-tuning **BART** with TensorFlow, achieving a ROUGE-L score of **0.82** and providing a comprehensive analysis of summarization techniques.
- Engineered and deployed the dual-methodology summarization engine as a real-time, interactive web application using **Flask**, providing an accessible interface for on-demand knowledge extraction and comparative analysis.

Emotion based music recommendation system — Flask, TensorFlow, RNN, RNN, BM25, LSA

- Engineered a novel emotion-aware music recommendation engine, leveraging a fine-tuned **LSTM-based deep learning model** to classify emotional sentiment from lyrical data with a **56% increase in prediction accuracy**.
- Designed a hybrid content-matching algorithm, integrating classical information retrieval (BM25) with **Latent Semantic Analysis (LSA)** to model thematic relevance, boosting overall recommendation quality by **36%**.
- Built and deployed a full-stack recommendation platform, using **Flask** for the interactive front-end and a **SQL** backend to manage user data and enable dynamic, real-time generation of personalized, emotion-driven playlists.