SHYAM SUBRAMANIAN

Principal Data Scientist | Multi-modal Systems | AI Agents | Large-Scale Unstructured Data
● shyamsubramanian24@gmail.com ● linkedin.com/in/shyamsubramanian ● github.com/shyamsubramanian

ABOUT ME

Principal Data Scientist transforming large-scale unstructured language and visual data into structured representations for Multi-modal LLM and Agentic Systems to reason over complex domain-specific patterns

EDUCATION

MBA, University of the Cumberlands, Williamsburg, KY — Dec '24

MS in Data Science, Worcester Polytechnic Institute, Worcester, MA — May '20

BE in Computer Science and Engineering, Anna University, Chennai, India — May '16

WORK EXPERIENCE

Fidelity Investments, Boston, MA

- **Principal Data Scientist** — Jan '24 – Present
- **Senior Data Scientist | Data Scientist** — Jun '20 – Jan '24

Agentic AI for Customer Service Request Automation

- Developed an Agentic AI system **(retrieval, tool calling, reflection, human-in-the-loop)** to automate classification, generation, and routing of customer service requests.
- Productionized on AWS EKS using LangSmith, integrating LLaMa and GPT models for scalable inference.
- Led pilot A/B testing, observing **~2 min time savings** per request and **6% reduction in rejection rate**, with projected impact on resolution time measures across **~400K requests annually**.

Call Intelligence for Contact Center Assistants (R&D)

- Curated and structured massive-scale call data (~18M calls over 1.5 years) from Snowflake, sampling, filtering, and complexity-stratifying calls to create high-quality datasets for modeling and evaluation.
- Designed curriculum-style **pre-training and fine-tuning for foundation models** on regulatory financial knowledge, internal policies/docs, and call datasets enabling models to capture domain-specific patterns.
- Constructed hierarchical **conversational tree datasets** at conversation and topic/subtopic levels, to support structured dialogue management and multi-turn interactions beyond intent-based rule systems.
- Developed and evaluated a **multi-turn Conversational RAG system** using the trained LLMs and conversational tree data, achieving 0.7 retrieval MRR, 4.6/5 average human rating, and 79% response acceptability during a 10-week labeling cohort with 20 phone representatives.

Data Science Core & Common Applications

- Built a custom **multi-modal document annotation tool**, streamlining data collection and ensuring high-quality inputs for downstream document processing systems.
- Developed and deployed an **end-to-end AI Model inventory management** system on AWS EKS with Jenkins CI/CD, integrating model risk, governance, and deployment workflows.
- Validated and established standards for Data Science accelerators, including Web apps, Labeling tools, Agentic AI frameworks, and LLM benchmarks.

Multi-Modal Document Understanding for 401-K Client Onboarding

- Developed and deployed automated information extraction pipelines from 401-K business documents, achieving **~85% accuracy across 80+ fields**, saving thousands of manual hours (~6 full-time employees)
- Engineered **multi-modal document parsing** algorithms to extract text, tables, checkboxes, and key-value pairs from complex, varied layouts, enabling structured representation of heterogenous documents.
- Implemented a custom retrieval and ranking system by **fine-tuning Bi-Encoder and Cross-Encoder** Sentence Transformer models, optimizing passage retrieval within documents.
- **Fine-tuned pre-trained and instruction-tuned LLMs** (BERT, T5, Flan-T5++) using **LoRA** for answer generation, enhancing model performance on complex documents for scalable, accurate information retrieval.

Data Science Intern, Fidelity Investments, Boston, MA — Jun '19 – Aug '19

- Developed a reusable PyTorch framework for **neural network-based time-series forecasting**, integrating models including DeepAR (Amazon) for multi-variate quantile regression analysis.
- Engineered scalable, multi-GPU deployment on Kubernetes using **Kubeflow and Docker**, abstracting GPU access for seamless model training and workflow orchestration.

Web Developer, Worcester Polytechnic Institute, Worcester, MA Jan ’19 – May ‘20

- Developed responsive, accessible web components and UI/UX features using **JavaScript, HTML/CSS, PHP**, and Drupal CMS, streamlining content management workflows.
- Improved website performance and accessibility compliance to achieve **~40% higher Google Lighthouse scores;** optimized database queries for faster load times and managed content to ensure up-to-date information.

Member Technical Staff, ZOHO Corporation, Chennai, India Jan ’16 – Jun ‘18

- Developed and deployed a high-performant web-based remote desktop control application for real-time IT troubleshooting with **C++, Java/JSP, and ReactJS**.
- Enabled low-latency, multi-monitor remote troubleshooting using websockets and multi-threading, facilitating over **185K+ monthly sessions** for 1,000+ high-value customers out of 280K+ global client enterprises.
- Integrated a text, voice, and video chat tool using WebRTC protocol with STUN and TURN server components with C++, Java/JSP, and ReactJS.
- Enabled peer-to-peer, multi-participant voice and video chat powered remote troubleshooting facilitating over **40K+ monthly remote sessions** for 1,000+ high-value customers.
- Developed a product level search functionality using AngularJS and Apache Lucene with additional natural language query matching enabling improved discoverability of the product.

Software Engineer Intern, ZOHO Corporation, Chennai, India May ’15 – Jul ‘15

- Developed IT tools for remote system inspection and help-desk support on Windows platforms with **C#, .NET** enabling remote system visibility and rich help-desk issue reporting.

PUBLICATIONS & PATENTS

Systems and methods for detection and extraction of borderless checkbox tables – *Granted in 2025* Aug ‘22

- Developed a robust algorithm to detect and extract borderless tables with checkboxes from business documents found predominantly in financial and legal domain
- Employed a visual grid-based segmentation and textual pattern-based expansion in OpenCV to approximate the fuzzy header, column, and row boundaries with text overlap between adjacent cells in the table

Systems and methods for automated end-to-end text extraction of electronic documents – *Granted in 2025* Jun ‘22

- Developed an empirically driven decision flow algorithm to combine native PDF text and OCR text for highly accurate text extraction from PDF documents
- Employed lookup-based automated text accuracy calculation and BERT model-based text auto correction to increase text extraction accuracy and decrease OCR latency

Hierarchical Evidence Set Modeling for Automated Fact Extraction and Verification – *EMNLP Publication* Oct ‘20

- Designed a novel multi-hop reasoning approach to construct evidence sets from large document collections, enabling robust fact extraction for claim verification.
- Developed a neural network, with novel aggregation components (contextual and non-contextual) aggregations for combining the evidence sets with hierarchical attention for claim verification
- Improved the state-of-the-art fact extraction and claim verification accuracy by 1.49% on FEVER dataset using language models (BERT, RoBERTa and ALBERT)

TECHNICAL SKILLS

Languages	Python, C++, Java, C#, JavaScript
Databases	Snowflake, DuckDB, Neo4J, MongoDB, PostgreSQL, MySQL, MSSQL
Big Data & Distributed Systems	Spark, Dask, Hadoop, Pig
Vector Databases	FAISS, Milvus, AWS OpenSearch
Deep Learning & LLMs	PyTorch, Tensorflow, Transformers, LangChain, LangGraph, LlamaIndex, vLLM, Ollama
Graph & Structured Data	Knowledge Graphs, Document Graphs, Conversational Trees, GraphRAG, Graph Neural Networks (GNNs), Think-on-Graph, Graph-of-Thoughts
MLOps & Deployment	AWS EKS, AWS Sagemaker, AWS Bedrock, Jenkins
Web Development	ReactJS, AngularJS, NodeJS, Django, Flask, FastAPI Streamlit, Reflex, Gradio