Devesh Maheshwari

Madison, WI 53703 | (608)-440-5616 | deveshmaheshwari25@gmail.com | linkedin.com/in/devesh32 |
github.com/Devesh-Maheshwari

Professional Summary

Seasoned AI Engineer with 4+ years of experience delivering Generative AI and NLP solutions in production. Specialised in end-to-end LLM application development, from RAG architectures to scalable vector search pipelines, with a strong record of enhancing model accuracy, efficiency, and business impact.

Education

University of Wisconsin–Madison *Madison, USA*
M.S. in Computer Science *Aug 2025 – Dec 2026*
Courses: Foundation Models, Advanced Natural Language Processing, Machine Learning, Data Exploration in Data Science, Mathematical Principal of RL, Responsible AI, Independent Research, Distributed System, Topics in DBMS, Operating Systems

Indian Institute of Technology (BHU) *Varanasi, India*
B.Tech. in Electrical Engineering *Aug 2017 – May 2021*
Courses: Data Structures and Algorithms, NLP, Probability and Statistics, Operations Research CGPA: 8.5/10

Work Experience

University of Wisconsin–Madison *Madison, USA*
Teaching Assistant — CS 320: Data Science Programming II *Spring 2026*

- Led practical lab sessions covering data preprocessing, feature extraction, and debugging of AI-adjacent Python workflows.
- Built automated validation and testing workflows to detect logical errors and edge cases, mirroring real-world challenges in AI.

Auxia *Bengaluru, India*
AI Research Engineer *Jan 2025 – Aug 2025*

- Built two standalone treatment embedding pipelines (**OpenAI + SBERT**) with enriched metadata, enabling reusable and client-specific representations and improving downstream target prediction accuracy by 12%.
- Automated and standardized ML training workflows with pipeline orchestration tools, reducing manual coding by 50%, cutting training time by ∼25%, and enabling scalable, cron-scheduled production deployments.
- Prototyped behavioral segmentation using LLM-generated user summaries from event logs successfully uncovering latent user groups—boosting CTR prediction accuracy by ∼9%.
- Led uplift modeling and A/B testing using funnel/binning analysis to identify high-lift segments and applied suppression/force-send rules to optimize targeting, achieving a 3.7× lift in purchase likelihood.

HiLabs *Pune, India*
Senior AI Developer *Apr 2023 – Jan 2025*

- Leveraged diverse generative and encoder-based LLMs (**T5, BioBERT+CRF/LSTM, Mistral, LLaMA**) for clinical entity recognition from medical charts, achieving a 95% F1 score in structured extraction of drugs and diseases.
- Optimised inference latency and deployment efficiency via model distillation, layer freezing and mixed-precision inference, reducing end-to-end extraction time while preserving entity-level precision for safety-critical medical information.
- Fine-tuned **T5** for clinical abbreviation expansion (93% accuracy).
- Led a team of 6 engineers leveraging **LLMs** for entity extraction from medical charts and Google search results which was later used in downstream RAG-based QA systems for healthcare clients.
- Designed a multi-class webpage classifier (93% precision) using **BERT** embeddings with XGBoost.
- Spearheaded the Provider Directory Accuracy product to improve HEDIS scores for U.S. health plans by validating provider directory data and recommending corrections for outdated records using multiple trusted data sources, driving Fortune 500 adoption and increasing ARR from $11M to $39M.
- Awarded the **Xtra Miler Award** for developing a critical data augmentation module that can provide multi-address recommendations.

Publicis Sapient *Gurugram, India*
AI Developer *Jun 2021 – Apr 2023*

- Converted multiple NLP models into containerized microservices deployed on AWS (**ECS/EKS)**, improving scalability and reducing cloud costs by 30%.
- Built a video processing pipeline to ingest media, generate transcripts, and create searchable MCQs, reducing manual content preparation time by 70%.
- Designed AI-enabled chatbot for a financial client, saving 2800+ man-hours, reducing support needs by 80%, and cutting case volumes by 4300+ per quarter; received a **Certificate of Appreciation**.

Projects

Multi-Agent System for Automated Market Report Generation

- Designed a LangGraph-based multi-agent workflow (researcher, writer, critic) with explicit task decomposition and dependency tracking, automating end-to-end report generation and reducing analyst effort by **8 hours/report** across repeated runs.
- Implemented a critic-driven self-reflection loop with source verification and argument consistency checks, reducing factual errors by **41%** and improving average user rating from 3.2 to **4.5/5** in internal evaluations.
- Integrated tool-augmented generation (web search, SQL analytics, charting) with model-agnostic interfaces, producing 10-page reports in **5 minutes** with reproducible outputs and traceable sources.

LLM-Powered Data Extraction from Unstructured Documents

- Fine-tuned Llama-3-8B to extract structured JSON from messy invoices/receipts; **94% field accuracy** vs 78% with GPT-4.
- Processing pipeline handles 7 document formats (PDF, scans, emails); deployed on GCP with **10K docs/day** throughput

Publication & Certifications

- *Beyond Self-Refinement: Ensembling and Chaining for Neurosymbolic Reasoning*, ICLR 2026
- *An Empirical Evaluation of Machine Learning for Hardening Security Devices in Data Networks*, IEEE Chilean,2021

LLM Engineering Azure Fundamentals (AZ-900) Azure AI Fundamentals (AI-900)

Skills & Interests

Programming Languages: C, C++, Python, PySpark, Bash
NLP: Transformers, BERT, GPT, Generative AI, Mamba, RAG, LSTM, RNN
Tools: Airflow, ElasticSearch, Celery, RabbitMQ, Docker, Kubernetes, Prometheus, Grafana, LangSmith
Frameworks: Django, PyTorch, TensorFlow, LangGraph, LangChain
Databases: MySQL, MongoDB, Chroma, Bigtable, BigQuery, Snowflake, PostgreSQL, Redis
Cloud Services: AWS, GCP
Interests: Generative AI, NLP, Deep Learning, Machine Learning, Data Science, Competitive Programming, Problem Solving