

# SAI AKSHAY MENTA

+1 (617) 406-7914 | [menta.sa@northeastern.edu](mailto:menta.sa@northeastern.edu) | Boston, MA | [linkedin.com/in/akshay-menta](https://www.linkedin.com/in/akshay-menta) | [akshay-menta.github.io](https://akshay-menta.github.io)

## EDUCATION

**Northeastern University - Boston, MA**  
Master of Science in Artificial Intelligence  
**Relevant Coursework:** Foundations of Artificial Intelligence, Natural Language Processing, Large Language Models, Machine Learning, Advanced Perception, Algorithms, Program Design Paradigm

**Sep 2024 - May 2026 (Expected)**  
GPA: 3.7/4.0

## EXPERIENCE

**Data Science Intern**  
Genpact

**Feb 2024 - Jul 2024**  
Hyderabad, India

- Engineered an end-to-end **RAG (Retrieval-Augmented Generation)** system using **LangChain** and **Pinecone** to automate enterprise knowledge retrieval, reducing manual query time by **30%**.
- Optimized internal analytics workflows by implementing semantic search on unstructured data, improving response accuracy by **25%** and streamlining the document review process for business stakeholders.
- Developed automated data visualization pipelines in **Power BI** to track pricing KPIs, enabling data-driven decision-making that directly impacted digital marketing strategies.
- Collaborated within an **Agile** team (Jira), working alongside Data Engineers to align model requirements with data infrastructure and business logic.

**Research Assistant**  
Amrita School of AI

**Nov 2022 - Jan 2024**  
Coimbatore, India

- Architected a **multimodal RL** framework fusing vision-text embeddings (CLIP) to outperform PPO baselines; accepted at **ACL 2024** as part of 4 peer-reviewed publications (IEEE, Springer).
- Engineered a speech recognition pipeline by fine-tuning **Whisper Large** using **LoRA (PEFT)**, reducing computational overhead while boosting accuracy by **10%** on dysarthric speech data.
- Developed a medical diagnostic system achieving **84% accuracy** by implementing **Diffusion Models** for synthetic data generation and **Grad-CAM** for model interpretability.

## PUBLICATIONS

- Semi Supervised Flood Damage Detection Using Satellite Images**, Published in Springer Nature, ICCAIML 2024. doi: [10.1007/978-981-96-0451-7\\_11](https://doi.org/10.1007/978-981-96-0451-7_11)
- Enhancing Knee Osteoarthritis Severity Classification Using Diffusion Augmented Images**, Published in Springer ICACECS 2023. doi: [10.2991/978-94-6463-314-6\\_27](https://doi.org/10.2991/978-94-6463-314-6_27)
- Few Shot Dysarthric Speech Intelligibility Classification Using Transformers**, Published in IEEE ICCCNT 2023. doi: [10.1109/ICCCNT56998.2023.10308067](https://doi.org/10.1109/ICCCNT56998.2023.10308067)
- Text Guided Reinforcement Learning Using Dravidian Commands**, Published in ACL DravidianLangTech 2023. [ACL Anthology](#)

## PROJECTS

**WhisperSense: Semantic-Aware Audio Captioning**  
Tech Stack: Python, Whisper, ConvNeXt, CLAP, FENSE, PyTorch

**Sep 2025 – Dec 2025**

- Engineered a confidence-calibrated pipeline fusing **Whisper** and **ConvNeXt**, utilizing similarity-driven decoding to boost **CLAP score by 18%** and reduce calibration error (ECE) by **22%** across 7k+ audio samples.

**AutoQuest: Automated Q&A Generation (RAG)**  
Tech Stack: Python, T5-Small, FAISS, FastAPI, Docker, Kubernetes

**Jan 2025 – Apr 2025**

- Deployed a scalable Q&A microservice using fine-tuned **T5-Small** and **FAISS**, achieving **<800ms latency** on 50k+ chunks by orchestrating containerized inference with **Docker** and **Kubernetes**.

## SKILLS

|                      |                                                                                   |
|----------------------|-----------------------------------------------------------------------------------|
| Languages            | Python, SQL (PostgreSQL), C++, Java, Bash                                         |
| AI & GenAI           | PyTorch, Transformers (Hugging Face), LangChain, LlamaIndex, RAG, FAISS, Pinecone |
| Data & ML Libraries  | TensorFlow, Scikit-learn, XGBoost, Pandas, NumPy, OpenCV                          |
| Cloud & MLOps        | AWS (SageMaker, S3, EC2), Docker, Kubernetes, MLflow, CI/CD, Terraform            |
| Deployment & Backend | FastAPI, Flask, REST APIs, ONNX, Git, Linux                                       |