Vijendra Kumar Vaishya
Computer Science & Engineering
Indian Institute of Technology Bombay

24M2133
M.S. by Research
Gender: Male
DOB: 18/07/2001

Examination	University	Institute	Year	CPI / % Credits
Post Graduation	IIT Bombay	IIT Bombay	2027	8.9 **Core:** 28 **| Total:** 58
Graduation	RGPV	Jabalpur Engineering College	2024	8.06
Graduation Specialization: Information Technology				

RESEARCH & THESIS

- **Beyond Single-Axis Fairness: Learning to Detect Intersectional Biases - EACL 2027 (Under Review)**
  Authors: ***Vijendra Kumar Vaishya*** *, Nihar Ranjan Sahoo, Pushpak Bhattacharyya*
  - Built a **retrieval-based framework** for intersectional bias detection, recast as **retrieval over classification**.
  - Introduced **BiasRetriever**, a **contrastively trained retriever** using canonical single-axis stereotype embeddings.
  - Boosted **Jaccard score by 18 points** over the strongest LLM baseline across disjoint corpora.
- **MIST: Mamba-based Integrated Spatio-Temporal Model for Intracranial Vessel Segmentation in DSA Sequences - WACV 2027 (Under Review)**
  Authors: ***Vijendra Kumar Vaishya*** *, Viraj Singh Gaur, Amruth Sagar, Swapna Sourav Rout, Kshitij Sharad Jadhav*
  - Developed a **dual-branch spatio-temporal architecture** for DSA vessel segmentation using **Mamba-based SSMs**.
  - Designed a **Cross-Attention Fusion (CrossSTF)** module that integrates MinIP-derived spatial priors with multi-scale temporal features, refined using **topology-preserving deep supervision**.
  - Achieved a **Dice score of 0.7801** on the independent DSCA benchmark without any retraining or fine-tuning, demonstrating strong cross-dataset generalization across diverse acquisition settings and imaging protocols.
- **Ontology-Aware Persona-based Chatbot for Tobacco Cessation | KCDH Lab, IIT Bombay** *(April 2025 – Ongoing)*
  In collaboration with Public Health Foundation of India (PHFI)
  - Engineered a **2-stage conversational architecture** utilizing a **decision-tree knowledge graph** to dynamically classify patient readiness across **5 cessation stages** and assess Fagerström dependence severity.
  - Built a **6,900-concept clinical ontology** with **33k+ edges** and designed an ontology-aware text segmenter, reducing broken mid-sentence chunk splits from **84.6% to 5.6%** compared to fixed-size chunking.
  - Achieved **58% Recall@3 (29/50)** and **0.497 MRR** using ontology-aware chunking across 50 clinical queries, outperforming fixed-size and dynamic semantic baseline approaches.
- **Bhashini Speech-to-Speech Machine Translation (SSMT)** *(August 2023 – Ongoing)*
  (NLTM-Bhashini, MeitY, Govt. of India)
  - Architected a **full-stack annotation platform** and transformer labeling model for **Marathi Multi-Word Expression** detection on **4,260 manual annotations** to accelerate low-resource language corpus curation.
  - Engineered an **automated MLOps retraining engine**, expanding dataset size from **4,260 to 13,900+ annotations**.
  - Achieved **85.05% test F1** and **90.76% Complex Predicate detection**, surpassing baseline performance.
- **Bias Detection in LLMs Using RAG** *(Fall 2024)*
  (CS 777: R&D 1 Project | Guide: ***Prof. Pushpak Bhattacharyya****)*
  - Built a **RAG-based bias detection framework** using **Phi-3-mini** and FAISS to classify bias type from sentences.
  - Fine-tuned Phi-3-mini on 1,000 samples, achieving **60.4% accuracy**, a **10% gain** over the no-RAG baseline.

PRE-PRINTS

- **Subset Selection for Fine-Tuning: A Utility-Diversity Balanced Approach for Mathematical Domain Adaptation** *(Pre-print)*
  Kotecha M, ***Vaishya VK****, Gautam S, Racha S.*

COURSE PROJECTS

- **Medical Image Classification Using Vision Models** *(Fall 2024)*
  (CS 725: Foundations of Machine Learning | Instructor: ***Prof. Preethi Jyothi****)*
  - Built an AutoML optimized **Vision-Transformer** pipeline for **multi-modal** medical image classification.
  - Exhibited **6.7% average gain** in Accuracy and AUC on multiple **MedMNIST** datasets to **ResNet** and **DenseNet**.
- **Multi-Hop Question-Answering & Reasoning** *(Fall 2024)*
  (CS 626: Speech and Natural Language Processing and the Web | Instructor: ***Prof. Pushpak Bhattacharyya****)*
  - Implemented **self-prompted chain-of-thought** with **6 reasoning paths** and studied extensibility to **Indic** tasks.
  - Infusing **LightRAG** with GPT-4o, increasing EM by **2%** on HotpotQA and 2Wiki, compared to Zero-Shot CoT.
  - Experimented **NLP** techniques and CoT models on Indic languages through **iNLTK** for dataset pre-processing.
- **Subset Selection for Fine-Tuning LLMs: Utility-Diversity Balanced Approach** *(Spring 2025)*
  (CS769: Optimization in ML | Guide: ***Prof. Ganesh Ramakrishnan****)*

- Proposed a **utility-diversity subset selection framework** for LLM fine-tuning, introducing a **Chain-of-Thought (CoT) metric** combining confidence divergence, backtrack frequency, and complexity to score difficulty.
- Demonstrated the effectiveness of our **data selection method**, achieving **50% accuracy at 1900 examples**, outperforming DPP and random selection.

- **Diagram Visual Question Generation using LeJEPA and BLIP-2** *(Fall 2025)*
  *(CS 772: Deep Learning for Natural Language Processing | Instructor: **Prof. Pushpak Bhattacharyya**)*
  - **Engineered** a two-stage visual question generation pipeline combining Stage 1 **LeJEPA** pre-training with Stage 2 **BLIP-2 Q-Former** fine-tuning, **outperforming** the baseline by **4 BLEU-4** points.
  - **Surpassed** 7.8B SOTA model on **DiagramQG** in **Bert-Score** (93.55 vs 90.98) and **METEOR** (51.75 vs 50.73) by coupling a **LeJEPA** encoder with **FLAN-T5-Large**, **reducing** parameters by **5.5x**.
- **Controlled LLM Decoding via Sequential Sampling Methods** *(Fall 2025)*
  *(CS 791: Probabilistic Foundations of AI | Instructor: **Prof. Sunita Sarawagi**)*
  - **Evaluated** 7 controlled decoding strategies including **SMC and TSMC**, achieving a **146% increase in target rarity reward** over baseline greedy decoding while maintaining low perplexity.
  - **Built** a custom evaluation benchmark across **20 test prompts**, demonstrating that SMC achieves **32% higher sample diversity** than SIS while eliminating particle weight degeneracy.
- **AQI Guardian – Real-Time Air Quality Platform** *(Fall 2024)*
  *(CS 699: Software Lab | Instructor: **Om P. Damani**)*
  - Built a **Flask** backend integrating weather APIs to serve real-time pollutant metrics for global metro cities.
  - Built a **BeautifulSoup** scraper mapping pollution levels to health gear, bypassing CAPTCHAs and rate-limits.

SELF PROJECTS

- **Fused SwiGLU / Transformer FFN Inference Engine**
  - Built **bare-metal C++17 engine** for LLaMA/Mistral FFN with a **5-stage optimization ladder** to **INT8**.
  - Wrote **8-wide NEON microkernel**, reaching **37–84x speedup** over scalar baseline.
  - Applied **INT8 quantization** to cut compute cost, hitting **41.7 GFLOP/s** with under 1% error.
  - Fused operations into **CPU registers** to cut memory traffic, boosting throughput **1.43–1.48x** on Apple, Snapdragon.
- **Custom Gaussian Process & Bayesian Optimization Engine**
  - Implemented **from-scratch GP surrogate** using Cholesky factorizations (LL^T) and log-marginal likelihood optimization.
  - Formulated **EI & PI acquisition functions** via analytical Gaussian CDF/PDF, converging in 50 evaluations.
- **StoriesbySafarnama - Collaborative Digital Storytelling Platform**
  - Founded and built **Safarnama**, a private platform for creating and sharing collaborative **digital memory timelines**.
  - Designed and deployed **full-stack architecture** for timeline creation and collaborative editing, integrating media uploads and secure sharing. Built scalable backend to manage **user data, timeline content, and media assets**.
  - Integrated **DODO Payments** for secure checkout, transaction tracking, and automated refund workflows.

COURSES TAKEN

- Foundations of Machine Learning
- Deep Learning for Natural Language Processing
- Speech and Natural Language Processing and the Web
- Software Lab
- Optimization in Machine Learning
- Probabilistic Foundations of AI
- Organization of Web Information

POSITIONS OF RESPONSIBILITY

- **Student Companion** | **ISCP, IIT Bombay**
  - Part of a **330-member team** selected on the basis of comprehensive interviews.
  - Mentored **5 students** in their academics, social, and extracurricular activities, ensuring their holistic well-being.
- **Technical Team Member** | **AACL 2025, IIT Bombay**
  - Supported technical execution of a **5-day international conference**, ensuring smooth session operations.
  - Managed **technical operations across venues**, troubleshooting issues to ensure uninterrupted sessions.

TECHNICAL SKILLS

- **Languages:** Python, C++ (17), C, Bash Scripting
- **ML/DL Frameworks:** PyTorch, Hugging Face Transformers, OpenCV, NumPy, CUDA
- **ML Techniques:** Mamba/State-Space Models, Vision Transformers, RAG, Contrastive Learning, LLM Fine-tuning & Quantization (INT8), Bayesian Optimization
- **Systems & Backend:** ARM NEON SIMD, Flask, MLOps Pipelines, FAISS, Docker, MySQL
- **Tools:** Git, LaTeX, Postman

ACHIEVEMENTS & EXTRA-CURRICULAR ACTIVITIES

- Secured **3rd District Rank** in Class 12th Board Examinations; qualified **JEE Advanced**.
- Scored **100%** in Mathematics in **CBSE** Boards (2018); solved over **500+ problems** on **LeetCode**.
- **Hobbies:** Gym, reading, Badminton, cooking, Cricket, Expressive Drawing, Chess & currently learning to play the guitar.