Vishvesh Trivedi

vishvesh106@gmail.com www.linkedin.com/in/vntrivedi www.github.com/NerdyVisky

US Permanent Resident (Green Card) — No Visa Sponsorship required

Education

New York University, Courant Institute of Mathematical Sciences **Sep 2024 – May 2026**
Master of Science, Computer Science, *GPA: 4.0/4.0*

National Institute of Technology, Surat **Sep 2020 – Jan 2024**
Bachelor of Technology, Computer Science, *GPA: 9.07/10.0*

Coursework: Data Structures & Algorithms, Programming Languages, Deep Learning, CV, NLP, Operating Systems, Database Management, Computer Architecture, Networks, Probability & Statistics, Mathematics, GPU

Selected Publications

C=Conference, J=Journal, S=In Submission

[C.1] S. Maniyar*, **V. Trivedi***, A. Mondal, A. Mishra, and C.V. Jawahar (2025). **AI-Generated Lecture Slides for Improving Slide Element Detection and Retrieval**. *ICDAR 2025 (ORAL, Top 2%).*

Experience

CILVR Lab, New York University **May 2025 – Present**
Graduate Research Assistant *New York, United States*

- Working with Prof. Eunsol Choi on optimizing inference and accuracy of in-context fact retrieval in multilingual LLMs.
- Modifying attention mechanisms of LLaMa-3.2-8B, Qwen-2.5-7B-Instruct, and Phi-3.5-3B-Ministruct on 5 different languages to improve factual retrieval by 15% compared to strong English baselines and 30% drop in KV-cache budget

Biomedical Data Sciences Hub, NYU Langone Health **Nov 2024 – Present**
Machine Learning Engineer *New York, United States*

- Developing production-grade safety modules that monitor ML models used in clinical workflows across 23 hospitals.
- Designed automated model drift detection pipelines using statistical tests (K–S, PSI, DeLong) to flag real-time deviations in model behavior and continuous post-deployment model monitoring using Prometheus and Grafana.
- Experience with HIPAA-compliant datasets like EPIC COSMOS, OMOP CDM of over 300M US patients using advanced SQL. Certifications in Caboodle, OHDSI and Clarity Data Models.
- Contributed to NIH and PCORi grant proposals on health-AI Safety and winner at the NYU Threesis Challenge Video

Center for Visual Information Technology, IIIT Hyderabad **Jan 2024 – Aug 2024**
Machine Learning Researcher *Hyderabad, India*

- Orchestrated a novel LLM-based pipeline to generate 18,000 high fidelity synthetic slides using university textbooks.
- Trained VLMs like LayoutLMv3, LLaVa-1.5-13B, CLIP on synthetic data to gain performance on Slide Element Detection and Retrieval tasks by 13% mAP and 10% Recall@K respectively, surpassing then SOTA benchmarks.
- Published findings as an oral presentation at ICDAR 2025. Over 2000+ visits, 500 downloads on HuggingFace. Website

Wells Fargo **May 2023 – Jul 2023**
Software Development Engineer Intern *Hyderabad, India*

- Pioneered a web-based fullstack tool using React and Typescript that produces semantic-aware audio-transcriptions of PPT presentations that is 40% faster than screen-readers, and directly impacts 15000 visually impaired WF employees.
- Spearheaded the team showcase event and completed 4 professional certifications on ML-Ops best practices, Data Governance, Agile, and Scrum methodologies. Received full-time return offer but turned down for higher studies.

Projects

Attention-Aware DPO for Reducing Hallucinations in Multi-Image QA [Code] [Website] [Report]
Hugging Face, PyTorch, Python, Bash, HPC, LLM-as-a-judge, Machine Learning, Deep Learning

- Trained LLaVa-1.5 with a novel Attention DPO loss function to increase multi-image VQA accuracy by 8.5%
- Used AdaptVis to optimize model performance at inference and push performance gain to 10% over base model.
- Devised a powerful LLM-as-a-judge using Gemini-2.5-Pro to rate outputs on quantifiable heuristics.

Open Source contribution to Retrieval Heads project [Code]
vLLM, ZeRO, flash-attention, PyTorch, Python, Hugging Face Transformers, GitHub, Open Source

- Rewrote the codebase of Retrieval Heads (ICLR 2025 spotlight paper) to make it run faster and consume less memory
- Designed high-throughput dynamic dataloaders in Pytorch, vectorized all tensor operations, and used flash-attention library and vLLM framework to bring down inference time by ×4 times (from 2hrs to 30mins) per experimental run.

Technical Skills

Languages: R/Python, C/C++, Java, SQL (Postgres, MySQL), XML, HTML/CSS, JavaScript, TypeScript, Bash/Zsh
Tools/Technologies: AWS, React, REST APIs, , GCP, Azure, Databricks, Docker, GIT, Mongodb, Redis
Frameworks: Sklearn, Pandas, Numpy, Pytorch, TensorFlow, Matplotlib, LangChain, Django, Streamlit