Avani Gupta

avani17101.github.io | avani17101@gmail.com | Google Scholar | GitHub: avani17101

Education

IIIT Hyderabad Hyderabad, India
B.Tech (Honours) + Masters By Research in Computer Science & Engineering, 2022
Advisor: Prof. P. J. Narayanan, MS specialization: AI, MS CGPA: 9/10
Thesis: Leveraging Human-Centered Explanations for Model Improvement & Evaluation Link

Peer-Reviewed Publications

- **Prototype Guided Backdoor Defense via Activation Space Manipulation**†, ICCV 2025.
  Venkat Adithya Amula, Sunayana Samavedam, Saurabh Saini, **Avani Gupta**, P. J. Narayanan.
  [paper] | [project page]
- **CAV Styler: Interpretable & Controllable Style Transfer**†, To appear in ICVGIP 2025.
  Sunayana Samavedam, Venkat Adithya Amula, Saurabh Saini, **Avani Gupta**, P. J. Narayanan.
- **Concept Distillation: Leveraging Human-Centered Explanations for Model Improvement** NeurIPS 2023. **Avani Gupta**, Saurabh Saini, P. J. Narayanan. [paper] | [project page]
- **Interpreting Intrinsic Image Decomposition using Concept Activations**, ICVGIP 2022 **(Oral + Best Paper)**. **Avani Gupta**, Saurabh Saini, P. J. Narayanan. [paper] | [project page]
- **Building Trust in Clinical LLMs: Bias Analysis & Dataset Transparency**, EMNLP 2025.
  Svetlana Maslenkova, Clement Christophe, Marco AF Pimentel, Tathagata Raha, Muhammad Umar Salman, Ahmed Al Mahrooqi, **Avani Gupta**, Shadab Khan, Ronnie Rajan, Praveenkumar Kanithi. [paper]
- **Med42 - Evaluating Fine-Tuning Strategies for Medical LLMs: Full-Parameter vs. Parameter-Efficient Approaches**, AAAI FM Symposium 2024.
  Clement Christophe, Praveenkumar Kanithi, Prateek Munjal, Tathagata Raha, Nasir Hayat, Ronnie Rajan, Ahmed Al Mahrooqi, **Avani Gupta**, Muhammad Umar Salman, Marco AF Pimentel, Shadab Khan, Boulbaba Ben Amor. [paper] [HuggingFace]
- **Predicting Business Process Events in Presence of Anomalous IT Events**, CODS-COMAD 2024. **Avani Gupta**, Avirup Saha, Sambit Ghosh, Neelamadhav Gantayat, Renuka Sindhgatta. [paper]
- **CitRet: A Hybrid Model for Cited Text Span Retrieval**, COLING 2022.
  Amit Pandey*, **Avani Gupta***, Vikram Pudi. [paper]

Preprints

- **A survey on Concept-based Approaches For Model Improvement***, **Avani Gupta**, P. J. Narayanan. arXiv:2403-14566
- **Goal-Oriented Next Best Activity Recommendation**,Prerna Agarwal*, **Avani Gupta***, Renuka Sindhgatta, Sampath Dechu. arXiv:2305.03219

Patent

- **Generating organizational goal-oriented & process-conformant recommendation models using artificial intelligence techniques**, [US Patent Application], Status: Pending. Prerna Agarwal, Sampath Dechu, **Avani Gupta**, Renuka Sindhgatta

Research Experience

MBZUAI AI Engineer, Research Office
Abu Dhabi, UAE Jun 2025 - Present

- Translated 10+ real-world industry problems into research formulations & identified relevant faculty collaborators, combined academic grounding with practical feasibility.
- Proposed a new MARL training paradigm aimed at fostering peer learning & survival instincts in AI Models.
- Designed novel peer-learning loss & survival-shaped rewards with promising preliminary results (ongoing).
- Built an Email Assistant agent (prioritization, extraction, drafting, calendar sync) & automated news-generation agents for the Research Office Newsletter (2000+ recipients).

*Equal Contribution †Mentored student-led work

- Designed LLM-as-judge pipelines for QA correctness, safety, bias evaluations & persona-conditioned synthetic data generation pipelines enabling scalable multi-persona evaluation.

Stealth AI Startup **AI Engineer**

Abu Dhabi, UAE Apr 2024 - Jun 2025

- Trained small LLMs (1-3B parameters, 2T tokens) for efficient edge deployment; proposed a novel MLA+GQA hybrid attention mechanism improving long-context efficiency for coding & tool-centric tasks.
- Built production-scale agents serving 3,000+ users (RAG, SQL, voice-call) using MCP/A2A protocols.
- Developed custom safety stack: content moderation model (self-trained), jailbreak detection, entity anonymization, & hallucination filters.

G42 Healthcare **AI Research Associate**

Abu Dhabi, UAE Mar 2023 - Apr 2024

- Modelled large-scale EHR sequences & trained a foundation model for patient trajectory prediction (2.2 million records), including chronic disease predictions, personalised medicine recommendation & mortality forecasting.
- Obtained patient embeddings revealing clinical coherence (chronic disease clusters, comorbidities).
- Orchestrated training datasets from 10M+ articles & contributed evaluation pipelines for **Med42** (70B open-source medical LLM achieving 72% on USMLE, AAAI FM Symposium 2024, EMNLP 2025).

Centre for Visual Information Technology (CVIT), IIIT Hyderabad **Researcher**

Hyderabad, India May 2020 - Mar 2023

- Originated a unified line of work on concept-driven representation learning across interpretability, vision, & graphics as part of my MS thesis.
- Introduced the concept-sensitivity loss (NeurIPS 23) & sensitivity-based disentanglement metric (ICVGIP 22).
- Guided student-led extensions in backdoor defense (ICCV 25) & style transfer (ICVGIP 25).

IBM Research **Research Intern**

Bangalore, India Sep 2022 - Dec 2022

- Developed anomalous IT events prediction using transformers (CODS-COMAD 24).

IBM Research **Research Intern**

Bangalore, India May 2021 - Aug 2021

- Formulated RL-based next-best-action recommendation for business processes with conflicting objectives; used action masking to handle dynamically evolving action spaces from process flow graphs (Patent; deployed on IBM Business Automation Workflow).

Honors & Awards

- **Best Paper Award**, ICVGIP 2022 1/400+ papers. [Link]
- **Global Rank 14**, Amazon ML Challenge 2021 3,000+ teams, 12K participants.[Link]
- **National Winner**, Smart India Hackathon 2020 Top 0.1% from 100,000+ teams.[Link]
- **National Winner**, Microsoft Mars Program 2020 Selected from 100,000+ students. [Link]
- **Dean's List** (Top 1% SGPA, Monsoon 2021) & **Merit List** (Top 5% SGPA, Spring 2021), IIIT Hyderabad

Service & Leadership

- **Reviewer**: NeurIPS 2025; ACL 2024; AAAI 2023.
- **Teaching & Mentorship**: Delivered lectures on **NeRFs**, **GANs**, and **Interpretability** for the 3D Vision School and IIIT-H Computer Graphics courses (2021 - 2022); mentored professionals and conducted tutorials for TalentSprints AI program (2022 - 2023).
- **Workshops & Community Engagement**: Instructor for the **GANs Workshop** [GitHub]; organizer of the AIcrowd *ML Battleground* challenge (2021) [Link]; student volunteer, NeurIPS Deep RL Workshop 2021.
- **Student Organizations**: Events Wing Member, Robotics Club, IIIT Hyderabad.