

# Avani Gupta

[avani17101.github.io](https://avani17101.github.io) | [avani17101@gmail.com](mailto:avani17101@gmail.com) | [Google Scholar](#) | [GitHub: avani17101](#)

## Education

### IIIT Hyderabad

Hyderabad, India

B.Tech (Honours) + Masters By Research in Computer Science & Engineering, 2023

Advisor: Prof. P. J. Narayanan, MS specialization: AI, MS CGPA: 9/10

Thesis: Leveraging Human-Centered Explanations for Model Improvement & Evaluation [Link](#)

## Skills

**Languages/Libraries:** Python, C/C++, PyTorch, TensorFlow, JAX, transformers.

**Frameworks:** RLlib, MARL (Mujoco, Gym), Distributed Training (DDP, FSDP), Synthetic Dataset Generation (Distilabel), AI Agents Development (MCP, A2A protocols, LangChain, ADK), FastAPI, SQL.

## Research Experience

### MBZUAI AI Engineer, Research Office

Abu Dhabi, UAE

Jun 2025 - Present

- Translated 10+ industry problems into research formulations, facilitating faculty-industry partnerships; also built and presented use-case demos to visiting industry delegations which led to active collaborations and signed MOUs.
- Independently proposed a collective-survival-based multi-agent reinforcement learning paradigm; designed novel peer-learning loss & survival-shaped rewards with promising results on benchmarks like PettingZoo and Smac-v2 (research ongoing with [Prof. Abdalla](#) and [Prof. Razzak](#)).
- Built AI agents for Research Office operations (email assistant prototypes, automated news generation) and generated synthetic persona-conditioned datasets for evaluation of RAG based AI assistant for correctness and safety using LLM-as-judge.

### Stealth AI Company AI Engineer

Abu Dhabi, UAE

March 2024 - May 2025

- Trained small LLMs (1-3B parameters, 2T tokens) for efficient edge deployment (using FSDP, megatron); proposed a novel MLA+GQA hybrid attention mechanism improving compute efficiency for coding & tool-calling capabilities.
- Generated Synthetic Datasets (via Distilabel) and worked on Supervised Finetuning for code and tool calling.
- Built production-scale agents serving 3,000+ users (multi-modal RAG, SQL, voice-call, marketing content generation including social media posts and high quality posters) using MCP/A2A protocols; also built a custom safety stack: content moderation, jailbreak and toxicity detection and flagging, & hallucination filters.

### G42 Healthcare AI Research Associate

Abu Dhabi, UAE

Mar 2023 - Mar 2024

- Modeled large-scale EHR sequences & trained a foundation model for patient trajectory prediction (2.2 million records), including chronic disease predictions, personalized medicine recommendation & mortality forecasting.
- Obtained patient embeddings revealing clinical coherence (chronic disease clusters, comorbidities).
- Curated training corpora from 10M+ articles & contributed evaluation pipelines for [Med42](#) (70B open-source medical LLM achieving 72% on USMLE, AAAI FM Symposium 2024) and 89B dataset for building trust via bias analysis (EMNLP 2025).

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

Hyderabad, India

May 2020 - Mar 2023

- Originated a unified line of research under my advisor Prof. P. J. Narayanan on concept-based learning for inducing knowledge priors, interpretable control, and debiasing neural networks.
- Introduced the concept-sensitivity loss (NeurIPS 23) & sensitivity-based disentanglement metric (ICVGIP 22).
- Guided students in research on backdoor defense (ICCV 25) & controllable style transfer (ICVGIP 25).
- Also worked on neural rendering, particularly NeRFs and temporal consistency in loose clothing animations

of 3D Human with Prof. Avinash Sharma.

### IBM Research Research Intern

Bangalore, India

Sep 2022 - Dec 2022

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

### IBM Research Research Intern

Bangalore, India

May 2021 - Aug 2021

- Formulated reinforcement learning based next-best-action recommendation for business processes with conflicting goals; trained PPO with action masking to handle dynamically evolving action spaces from process flow graphs (deployed on IBM Business Automation Workflow, US patent application pending).

## Honors & Awards

---

- **Best Paper Award**, ICGVIP 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
- **Merit-Based Chief Minister's State Scholarship** (2017, 2018), Chhattisgarh

## Service & Leadership

---

- **Reviewer**: CVPR 2026; NeurIPS 2025; ACL 2024; AAAI 2023.
- **Teaching & Mentorship**: Teaching Assistant for *Statistical Methods in AI* (Spring 2022); delivered talks on NeRFs, GANs, and interpretability at the 3D Vision Summer School and IIIT-H Computer Graphics course (2021 – 2022); AI Tutor and Mentor for the TalentSprint Professional AI Program (2022 – 2023).
- **Workshops & Community Engagement**: Instructor for the GANs Workshop [\[GitHub\]](#); Organizer of the Alcrowd *ML Battleground* challenge (2021) [\[Link\]](#); Student Volunteer at NeurIPS RL Workshop (2021).
- **Student Organizations**: Events Wing Member, Robotics Club, IIIT Hyderabad.

## Peer-Reviewed Publications

---

- **Prototype Guided Backdoor Defense via Activation Space Manipulation**, ICCV 2025.  
Venkat Adithya Amula, Sunayana Samavedam, Saurabh Saini, **Avani Gupta**<sup>†</sup>, P. J. Narayanan. [\[paper\]](#) | [\[project page\]](#)
- **CAV Styler: Interpretable & Controllable Style Transfer**, To appear in ICGVIP 2025.  
Sunayana Samavedam, Venkat Adithya Amula, Saurabh Saini, **Avani Gupta**<sup>†</sup>, 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**, ICGVIP 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. [\[paper\]](#).

\*Equal Contribution    †Mentored student-led work

---

## Preprints

---

- **A survey on Concept-based Approaches For Model Improvement**, Avani Gupta, P. J. Narayanan. [arXiv:2403-14566](https://arxiv.org/abs/2403-14566)
- **Goal-Oriented Next Best Activity Recommendation using Reinforcement Learning**, Prerna Agarwal\*, Avani Gupta\*, Renuka Sindhgatta, Sampath Dechu. [arXiv:2305.03219](https://arxiv.org/abs/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