Arjun Balaji

ab6136@columbia.edu www.arjunbalaji.in ArjunBalaji79

EDUCATION

Columbia University, *Masters in Public Administration* Aug 2025 – Present | New York City
Concentration in Technology Policy and Innovation. VP of Research, Columbia AI Club.

R.V. College of Engineering, *BE in Electronics and Telecomm.* Dec 2021 – Jun 2025 | Bangalore
Coursework: Multivariable Calculus, Differential Equations and Numerical Methods, Machine Learning, Statistics and Probability for Data Science, Linear Algebra, Integral Transforms and Fourier Series, Deep Learning, Signal Processing with Machine Learning

PUBLICATIONS

AI Safety, Alignment & Interpretability

- "Predicting Chain-of-Thought Correctness from Trajectory Geometry." Accepted at **TMLR** (August, 2026) **Link:** https://openreview.net/forum?id=H9cBkEqVeY
- "Certified by Abstention: Distribution-Free Guarantees for Chain-of-Thought Verifiers at Small Calibration Budgets." **AISTATS (under review, 2027)** Link: https://arxiv.org/abs/2610.09541v1
- "ForkSCOPE: Charting the Agentic Garden of Forking Paths" (working paper) Accepted at **AgentEvalSci** (October, 2026) Link: https://arxiv.org/abs/2609.12438
- "What the Guard Misses, the Robot Executes: Implied Harm in VLA Instructions." Accepted at **CoRL SPAIS Workshop (October, 2026).** Link: https://arxiv.org/abs/2610.05818
- "Sparse Spectral Signatures of Reasoning: Model-Agnostic Verification via Sentence-Level Graph Signals." **ICLR 2026 Workshop on Logical Reasoning of Large Language Models** (*https://iclr.cc/virtual/2026/10017393*)
- "Encoding Values: Injecting Morality into Machines via Prompt-Conditioned Moral Frames." Accepted at **NeurIPS 2025 Creative AI Track** (*https://neurips.cc/virtual/2025/loc/san-diego/129264*)
- "BiasCheck: An Analytical Framework for Contextual Bias Detection in Text, Models and Databases." Accepted at **AAAI 2025 Workshop on AI for Public Missions**

Applied ML

• Supervisory Text versus Contagion Graphs for Emerging-Market Crisis Early Warning: A Leakage-Controlled Benchmark on Free Public Data. (working paper)

• "Verifying the Erdős–Gyárfás Conjecture up to 31 Vertices with SAT Modulo Symmetries.". Link: https://doi.org/10.5281/zenodo.21190438

• ProteoDockNet: A Deep-Learning GNN Architecture for Ligand Binding Affinity Prediction from SMILES across Human Liver, Kidney and Brain Proteins. **Elsevier Computational and Structural Biotechnology Reports, 2024 (Q1 Journal)**

• A Comprehensive AI-Driven Decision Support System for Colorectal Cancer: Integration of Mutation Data and Web Application Implementation. Presented at **ICCPSMLA-24**

RESEARCH EXPERIENCE

New York University Courant, ML4Good Lab, *Researcher* Sep 2026
Researching LLM reasoning and stability with Prof. Daniel Neill and Prof. Emily Black

Columbia University Irving Medical Centre, *Research Assistant* Mar 2026 – Present | New York

- ARG Lab: Developing Voice AI agents and LLM pipelines for large-scale audio transcription and scoring across aging & dementia research studies at Columbia's Butler Aging Center with the Aiello Research Group.
- CDM Lab: Building an AI-powered dental medicine platform for radiograph interpretation training at Columbia's College of Dental Medicine, using VLM pipelines to help students analyze pathology and critically evaluate AI diagnostic output.

Carnegie Mellon University, *Research Intern* Sep 2024 – Mar 2025 | Pittsburgh

- Developed Unsupervised Models for Brain MRI Surface Registration in a collaborative project with **Harvard Medical School** at the Xu Lab.
- Designed and Implemented efficient geometric deep learning techniques for deforming a mesh template on to the brain surface.

Nanyang Technological University, *Research Intern* Oct 2024 – Jun 2025 | Singapore

- Researched advanced segmentation for MRI and echocardiograms at **MVAIT** (Medical Vision and Artificial Intelligence Technologies) Lab under the guidance of Prof. Si Yong.
- Built a dynamic 4D MRI memory bank with slice-level adaptive quantization (8/16/32-bit) for compact storage and novel rapid frame retrieval techniques.
- Designed and benchmarked multi-objective loss functions for custom architecture (Dice, Wasserstein, Cross-Entropy, Normal Consistency, Tversky, etc.) for UNet/CNN cardiac segmentation.

Indian Institute of Science, *Research Intern* Oct 2024 – Dec 2024 | Bangalore

- Developed graph sampling techniques and anomaly detection algorithms for detecting money laundering in financial transactions in collaboration with the National Payments Corporation of India (**NPCI**) at the VISTA Lab.

Samsung Research Institute, *Research Intern* Aug 2023 – Aug 2024 | Bangalore

- Built compression pipelines for Vision Transformer families (DeiT, ViT-B/16, Swin, MobileViT) combining post-training and quantization-aware training across W8/A8, W4/A4, and sub-2-bit regimes, evaluated on ImageNet classification and COCO detection.
- Designed multi-stage knowledge-distillation schedules like logit, intermediate-feature, and attention-map distillation under a progressive teacher–student curriculum to recover accuracy at aggressive bit-widths where single-stage distillation collapses.
- Delivered compressed ViT variants reaching up to 8× parameter and memory reduction while preserving classification accuracy within margin of the full-precision baseline.

PROFESSIONAL EXPERIENCE

ImpactAI Foundry, *Co-Founder and Program Director* Jan 2026 – Present

- Co-founded a venture that builds safe AI for the last mile. It runs cohorts with social impact organizations, builds products and policy, and does AI safety and alignment research grounded in real deployments. Part of **Decoded Futures** (Tech:NYC Foundation) and backed by **Emergent Ventures India**.
- Ran the first cohort in Bengaluru (June–July 2026) with 7 organizations across education, healthcare, livelihoods and women's empowerment, including Akshaya Patra, Reap Benefit and U&I.
- Secured paid engagements, including a Kannada voice AI worker-safety assistant for a national agricultural workers' union and AI-driven policy research on expanding social security. Built and delivered bilingual WhatsApp training and monitoring bots for Baala.
- Press coverage in TheCSRUniverse, TechGraph and CSRBox. More at: impactaifoundry.com.

Incluya, *Founding AI Engineer* Apr 2025 – Jul 2025

- Designed a personalized financial wellness AI using LLMs that adapt to user traits and emotional triggers.
- Integrated memory layers to retain evolving user profiles and improve long-term personalization.

People+ai, *AI/ML Developer* Jun 2024 – Mar 2025 | Bangalore

- Developed an application using Google Maps API and GRIHA guidelines to dynamically evaluate proximity to essential services to improve green building planning which has been publicly released and shared to both the Indian Government's GRIHA and developers.
- Utilized Stable Diffusion with ControlNet to generate synthetic site map data, leveraging edge maps and layout conditions to create precise and diverse datasets for training AI models.

RECOGNITIONS AND GRANTS

Emergent Ventures India Grant, Mercatus Center at George Mason University (2026)
Awarded $25,000 to ImpactAI Foundry, which builds safe AI for the last mile with social impact organizations. The grant funds a field lab where these organizations build their own AI tools, and every deployment becomes an open evaluation of where models fail in low-resource, safety-critical settings.

Lossfunk AI Resident (Batch 2, formerly Turing's Dream)
6-week in-person residency at the Lossfunk AI lab, BiasCheck (AAAI 2025 Workshop paper + open-source library); research on graph representation learning.

SAAR × AIM Intelligence Compute Grant
Competitive compute award supporting the chain-of-thought geometry research line (TMLR and EMNLP submissions).