Pratinav Seth

Mobile: +91 XXXXXXXXXX
Email: seth.pratinav@gmail.com
Personal Website: https://pratinavseth.github.io/
Google Scholar: Pratinav Seth
Linkedin: linkedin.com/in/pratinav-seth
Github: https://github.com/pratinavseth

Professional Summary

Research scientist working across mechanistic interpretability, post-training alignment, and LLM safety, with growing focus on AI agents and evaluation. 48 papers overall (31 peer-reviewed, 210+ citations) at venues including ICML, ACL, WWW, MIDL, and Nature Scientific Reports.

Recent Work Experience

Lead Research Scientist Apr. 2026 – Present
Lexsi Labs, Lexsi.ai (Aurionpro Solutions Group) *Remote*

- **Post-Training & Alignment**: Lead the lab's post-training and alignment research, directing workstreams across interpretability, safety, model optimization, and agents; scaled AlignTune (37⋆ [26]) into a production-grade multi-backend ecosystem spanning multiple SFT/RL algorithms, model merging, and domain-specific alignment auditing for BFSI, legal, and healthcare; shared applied results in case studies on template-strict domain specialization [49] and wealth-management alignment with AlignTune [50].
- **Post-Training & Data Curation**: Building CuratorKIT (24⋆) [27], a provenance-grounded data-curation and synthetic-generation pipeline [36]; verifiable long-form synthetic document generation, with further work under review [37].
- **Post-Hoc Interpretability**: Built unified post-hoc explainability tooling (LRP, Integrated Gradients, DL-Backtrace), including LRP- and DLB-based analysis of model refusal behavior.
- **Mechanistic Interpretability**: Lead circuit-level interpretability: built CircuitKIT (14⋆) [28], a circuit discovery, evaluation, and application toolkit; co-led C-$\Delta\theta$, a circuit-restricted weight-arithmetic method for selective refusal (**ICML 2026 Workshop** [9]); and advised machine unlearning via circuit attribution [38], with further work on circuit faithfulness under review [39].
- **Safety & Steering**: Improve safety alignment across the model lifecycle: during training, post-training recovery and fixing, and at inference time. Co-built SafeTune (7⋆) [29], a unified library for auditing and repairing safety drift in fine-tuned LLMs; built an inference-time alignment-transfer method (**ICML 2026 Workshop** [10]); a position paper on the limits of behavioral safety assurance for governance [33]; audited compliance detectors and guard models [40]; and further work on safety-drift repair and safety-repair evaluation, under review [41, 42, 43].
- **Model Optimization**: Lead structured-pruning and quantization work for efficient model deployment.
- **AI & Coding Agents**: Explore agent and coding-agent architectures; early exploration of a self-improving post-training agent and autonomous-research (auto-research) agents.
- **Evaluation**: Built an internal LLM evaluation library.
- **Research-to-Product Integration**: Translate research into production: integrate interpretability, alignment, and safety tooling into the product stack.
- **Team & Delivery**: Lead the Model Science group (16 interns and 8 full-time researchers throughout this period across India and Paris), covering grants, hiring, roadmap, and overall planning. Team output spans **ICML 2026** workshop papers and submissions under review at A* venues (**NeurIPS**, **ACL**, and more).

Research Scientist Jul. 2025 – Mar. 2026
Lexsi Labs, Lexsi.ai (Aurionpro Solutions Group) [Previously: AryaXAI, Arya.ai] *Remote*

- **Model Science Group**: Led and scaled the lab's Model Science group across post-training alignment, safety, and interpretability into a core research team driving the lab's alignment and safety agenda.
- **AlignTune**: Built AlignTune, a modular toolkit for post-training alignment of LLMs spanning SFT, preference optimization, and safety methods (37⋆ [26]); formalized the *Interpretability as Alignment* framework (**EurIPS 2025 Workshop** [34]) as a guiding design principle for the team's alignment work.
- **Interpretability**: Led the lab's interpretability tooling: rearchitected DLBacktrace v2 on a `torch.export`-based graph-capture design to keep it model-agnostic across architectures, with CUDA acceleration for LLMs and MoEs (26⋆ [30]); drove actionable interpretability into model optimization through interpretability-aware pruning for medical imaging (**MICCAI Workshop 2025** [11]).
- **Tabular Foundation Models**: Led problem framing, early model design, and benchmarking for the ORION tabular foundation-model series: Orion-Bix (**WWW 2026** [1]) and Orion-MSP (**AITD Workshop @ EurIPS 2025** [12]), and drove its extension to regression.
- **TabTune**: Architected TabTune, an open-source toolkit for inference, fine-tuning, and regression [51] with tabular foundation models (116⋆, **WWW 2026** Demo [31]), shipped with an accompanying fine-tuning study [2]; later advised distillation, ensembling (**ICML 2026 Workshops** [13, 14, 15]), and credit-risk prediction (**FinDS @ ACM SIGMOD 2026, Oral** [16]) for health and enterprise deployment.
- **Model Compression**: Built an internal pruning and model-compression toolkit, bringing interpretability-guided compression into the production deployment pipeline.

- **Team & Mentorship**: Led the Model Science group (14 interns and 5 full-time researchers during this period across India and Paris) spanning tabular, alignment, and interpretability; worked on grants, hiring, and overall planning.
- **Talks & Posters**: Spotlight talk at **EurIPS 2025 Workshop**; poster at **MICCAI Workshop 2025**.

Research Scientist Jul. 2024 – Jun. 2025
AryaXAI Alignment Labs (Arya.ai, Aurionpro Solutions Group) *Remote / Mumbai, India*

- **Model-Agnostic Explainability**: Generalized DL-Backtrace into a model-agnostic explainability framework for diverse deep learning architectures (**IJCNN 2025** [3]).
- **XAI Evaluation & Compliance**: Built xai_evals (15⋆) [32], a benchmarking framework for post-hoc local explanation methods, and argued for reliable XAI metrics in compliance settings (**EurIPS 2025 Workshop** [35]); focused on evaluation in high-stakes domains including fraud detection.
- **Leadership & Mentorship**: As one of the initial research hires, helped set up the team's research culture; mentored 2 research interns; led recruitment for interns and full-time scientists (Paris and India); presented AryaXAI solutions at the 5th MLOps Conference.

Research Intern Jan. 2024 – Jun. 2024
Rolnick Lab, Mila - Quebec AI Institute, Montreal, Quebec *Remote / Montreal, Canada*

- **Project Details**: Bachelor's thesis project (first-author, **ICML 2025** & **ICLR 2025 Workshop** [4]): computer vision and deep learning for geospatial applications targeting climate change, specifically detecting abandoned oil and gas wells. Worked with domain experts to build a novel geospatial dataset and benchmark deep learning models.
- **Mentor**: Dr. David Rolnick (McGill University, Université de Montréal, Mila).

Computer Vision Research Intern (CR/RDT-2) Jun. 2023 – Oct. 2023
Bosch Corporate Research (Robert Bosch Research and Technology Center India) *Bangalore, India*

- **Project Details**: Developed a generative data-augmentation pipeline for safety-critical autonomous-driving perception at a production R&D center. Used latent diffusion models to synthesize hard-negative and misclassified samples, improving downstream CV fine-tuning on rare edge cases relevant to ADAS safety.
- **Mentor**: Mr. Koustav Mullick (CR/RDT-2), under the guidance of Dr. Amit Kale.

Highlighted Achievements

- **48 papers overall, 31 peer-reviewed** (210+ citations, h-index 8): main conferences/journals (ICML, ACL, WWW, MIDL, IJCNN, AAAI, Nature Scientific Reports), workshops and shared tasks (CVPR, NeurIPS, ICLR, MICCAI, EurIPS, ACL, SIGMOD), plus library and position papers, and several first-author papers under review.
- **AAAI Undergraduate Consortium**: Scholar (2023) → Mentor (2026); 1 of 11 undergraduates globally selected as Scholar; topic: *Model-Agnostic, Uncertainty-Aware Machine-Learning Metrics*.

Open-Source Libraries

Designed and built the initial release of each library below.

Released: TabTune (116⋆, WWW 2026) · AlignTune (37⋆) · DLBacktrace v2 (26⋆) · xai_evals (15⋆) · CircuitKIT (14⋆) · CuratorKIT (24⋆) · SafeTune (7⋆)
In Development: Multiple tools across interpretability, post-training alignment, and LLM safety.

Publications

Main Conference

[1] Mohamed Bouadi, **Pratinav Seth**, Aditya Tanna, and Vinay Kumar Sankarapu. Orion-bix: Bi-axial attention for tabular in-context learning. In ***The Web Conference (WWW) 2026, ACM***, 2026. URL: https://arxiv.org/abs/2512.00181.

[2] Aditya Tanna, **Pratinav Seth**, Mohamed Bouadi, and Vinay Kumar Sankarapu. Exploring fine-tuning for tabular foundation models. In ***The Web Conference (WWW) 2026, ACM***, 2026. URL: https://arxiv.org/abs/2601.09654.

[3] Vinay Kumar Sankarapu, Chintan Chitroda, Yashwardhan Rathore, Neeraj Kumar Singh, and **Pratinav Seth**. Dl-backtrace: A model agnostic explainability for any deep learning models. *(All authors contributed equally)* ***International Joint Conference on Neural Networks (IJCNN) 2025***, 2025. URL: https://ieeexplore.ieee.org/abstract/document/11228632/.

[4] **Pratinav Seth**, Michelle Lin, Brefo Yaw, Jade Boutot, Mary Kang, and David Rolnick. Alberta wells dataset: Pinpointing oil and gas wells from satellite imagery. ***International Conference on Machine Learning (ICML) 2025 & ClimateAI Workshop, ICLR 2025***. URL: https://arxiv.org/abs/2410.09032.

[5] Danush Khanna, **Pratinav Seth**, Sidhaarth Sredharan Murali, Aditya Kumar Guru, Siddharth Shukla, Tanuj Tyagi, Sandeep Chaurasia, and Kripabandhu Ghosh. Self-percept: Introspection improves large language models' detection of multi-person mental manipulation in conversations. ***ACL (Main) 2025 & NAACL-SRW Workshop, NAACL 2025***. URL: https://arxiv.org/abs/2505.20679.

[6] Siddhant Bharadwaj, **Pratinav Seth**, and Chandra Sekhar Seelamantula. Obscure to observe: A lesion-aware MAE for glaucoma detection from retinal context. In ***Medical Imaging with Deep Learning (MIDL) 2025 - Short Papers***, 2025. URL: https://openreview.net/forum?id=gqLXT8Edf3#discussion.

[7] Aditya Kasliwal, Ishaan Gakhar, Aryan Kamani, **Pratinav Seth**, and Sriya Rallabandi. Lamar: Laplacian pyramid for multimodal adaptive super resolution. ***Student Abstract and Poster Program, Thirty-Eight AAAI Conference on Artificial Intelligence, AAAI 2024***, February 2024. doi:10.1609/aaai.v38i21.30463.

[8] Aditya* Kasliwal, Ishaan* Gakhar, Aryan* Kamani, **Pratinav Seth***, and Ujjwal Verma. Laplacian reconstructive network for guided thermal super-resolution. *(* - equal contribution)* ***Nature Scientific Reports***, 16, Jan 2026. doi:10.1038/s41598-026-36027-x.

Peer-Reviewed Workshop

[9] Aditya Kasliwal*, **Pratinav Seth***, and Vinay Kumar Sankarapu. C-$\delta\theta$: Circuit-restricted weight arithmetic for selective refusal. In ***Mechanistic Interpretability Workshop, ICML 2026*** *(* - equal contribution)*, 2026. URL: https://arxiv.org/abs/2602.04521.

[10] Chirag Chawla, **Pratinav Seth**, and Vinay Kumar Sankarapu. ALIGNBEAM: Inference-time alignment transfer via cross-vocabulary logit mixing. In ***AI for Good Workshop, ICML 2026***, 2026. URL: https://arxiv.org/abs/2606.12342.

[11] Nikita Malik, **Pratinav Seth**, Neeraj Kumar Singh, Chintan Chitroda, and Vinay Kumar Sankarapu. Interpretability-aware pruning for efficient medical image analysis. ***The 1st MICCAI Workshop on Efficient Medical AI***, 2025. URL: https://arxiv.org/abs/2507.08330.

[12] Mohamed Bouadi, **Pratinav Seth**, Aditya Tanna, and Vinay Kumar Sankarapu. Orion-msp: Hierarchical multi-scale attention for tabular in-context learning. In ***AI for Tabular Data (AITD) Workshop, EurIPS 2025***, 2025. URL: https://arxiv.org/abs/2511.02818, arXiv:2511.02818.

[13] Aditya Tanna, Nassim Bouarour, Mohamed Bouadi, Vinay kumar Sankarapu, and **Pratinav Seth**. Pocket foundation models: Distilling tfms into cpu-ready gradient-boosted trees. In ***Foundation Models for Structured Data (FMSD) Workshop, ICML 2026***, 2026. URL: https://arxiv.org/abs/2605.18654.

[14] Aditya Tanna, Yash Jignesh Desai, **Pratinav Seth**, Mohamed Bouadi, Nassim Bouarour, and Vinay kumar Sankarapu. Ensembling tabular foundation models: A diversity ceiling and a calibration trap. In ***Foundation Models for Structured Data (FMSD) Workshop, ICML 2026***, 2026. URL: https://arxiv.org/abs/2605.18696.

[15] Aditya Tanna, Nassim Bouarour, Mohamed Bouadi, Vinay kumar Sankarapu, and **Pratinav Seth**. Distilling tabular foundation models for structured health data. In ***Structured Data for Health (SD4H) Workshop, ICML 2026***, 2026. **Best Paper Runner-Up, Spotlight**. URL: https://arxiv.org/abs/2605.18702.

[16] Aditya Tanna, Mitul Solanki, Mohamed Bouadi, Nassim Bouarour, **Pratinav Seth**, and Vinay Kumar Sankarapu. Data presentation over architecture: Resampling strategies for credit risk prediction with tabular foundation models. In ***FinDS Workshop @ ACM SIGMOD 2026 (Oral)***, 2026. URL: https://arxiv.org/abs/2605.18635, arXiv:2605.18635.

[17] Krish Didwania*, **Pratinav Seth***, Aditya Kasliwal, and Amit Agarwal. Agrillm: Harnessing transformers for farmer queries. *(* - Equal Contributions)* ; ***NLP4PI Workshop @ EMNLP 2024 & Undergraduate Consortium at KDD 2024***. URL: https://arxiv.org/abs/2407.04721, doi:10.18653/v1/2024.nlp4pi-1.16.

[18] **Pratinav Seth** and Abhilash K. Pai. Does the fairness of your pre-training hold up? examining the influence of pretraining techniques on skin tone bias in skin lesion classification. In ***IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops***, 2024. doi:10.1109/WACVW60836.2024.00067.

[19] Dyutit Mohanty, Aditya Kasliwal, Bharath Udapa, and **Pratinav Seth****. Sailing through spectra: Unveiling the potential of multi-spectral information in marine debris segmentation. *(** - Senior Author)* ; ***The Second Tiny Papers Track at ICLR 2024***. URL: https://openreview.net/pdf?id=tJPLJS97X4.

[20] Aditya Kasliwal, Sankarshanaa Sagaram, Laven Srivastava, **Pratinav Seth****, and Adil Khan**. Refuseg: Regularized multi-modal fusion for precise brain tumour segmentation. *(** - Senior Author)* ; ***9th Brain Lesion (BrainLes) Workshop, MICCAI 2023***. URL: https://arxiv.org/abs/2308.13883.

[21] Aditya Kasliwal, **Pratinav Seth**, Sriya Rallabandi, and Sanchit Singhal. Corefusion: Contrastive regularized fusion for guided thermal super-resolution. In ***2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)***, pages 507–514, 2023. URL: https://arxiv.org/abs/2304.01243, doi:10.1109/CVPRW59228.2023.00057.

[22] **Pratinav Seth*** and Mihir Agarwal*. Uncertainty-aware test-time augmented ensemble of berts for classification of common mental illnesses on social media posts. In ***The First Tiny Papers Track at ICLR 2023 (* - equal contribution)***, *Kigali, Rwanda*, 2023. URL: https://arxiv.org/abs/2304.04539.

[23] Akshat Bhandari*, Sriya Rallabandi*, Sanchit Singhal*, Aditya Kasliwal*, and **Pratinav Seth****. Performance evaluation of deep segmentation models for contrails detection. *(* - equal contribution; ** - Senior Author)* ***Tackling Climate Change with Machine Learning Workshop, NeurIPS 2022***, 2022. URL: https://arxiv.org/abs/2211.14851.

[24] **Pratinav Seth**, Adil Hamid Khan*, Ananya Gupta*, Saurabh Mishra*, and Akshat Bhandhari**. Uatta-ens: Uncertainty aware test time augmented ensemble for pirc diabetic retinopathy detection. *(* - equal contribution; ** - Senior Author)* ***Medical Imagery Meets NeurIPS Workshop, NeurIPS 2022***, 2022. URL: https://arxiv.org/abs/2211.03148.

[25] Kumud Lakara*, Akshat Bhandari*, **Pratinav Seth***, and Ujjwal Verma. Evaluating predictive uncertainty and robustness to distributional shift using real world data. *(* - equal contribution)* ***Bayesian Deep Learning Workshop, NeurIPS 2021***, 2021. URL: https://arxiv.org/abs/2111.04665.

Library & Software Papers

[26] Chirag Chawla*, Zera Lyngkhoi*, **Pratinav Seth***, Utsav Avaiya, Soham Bhattacharjee, Mykola Khandoga, Rui Yuan, and Vinay Kumar Sankarapu. Aligntune: Modular toolkit for post-training alignment of large language models. (* - equal contribution) Pre-Print, 2025. URL: https://arxiv.org/abs/2602.09621, arXiv:2602.09621.

[27] Soham Bhattacharjee, Karun Sharma, Vinay Kumar Sankarapu, and **Pratinav Seth**. Curatorkit : Data curation and synthetic data generation for llm post-training. Pre-Print, 2026. URL: https://arxiv.org/abs/2606.21631, arXiv:2606.21631.

[28] **Pratinav Seth***, Hem Gosalia*, Aditya Kasliwal*, and Vinay Kumar Sankarapu. Circuitkit : Circuit discovery, evaluation, and application toolkit for mechanistic interpretability. (* - equal contribution) Pre-Print, 2026. URL: https://arxiv.org/abs/2607.19317, arXiv:2607.19317.

[29] **Pratinav Seth***, Saisab Sadhu*, Anshul Kaushal*, and Vinay Kumar Sankarapu. Safetune: A unified faithful library for auditing and repairing safety drift in fine-tuned llms. (* - equal contribution) Pre-Print, 2026. URL: https://github.com/Lexsi-Labs/SafeTune.

[30] Neeraj Kumar Singh*, **Pratinav Seth***, Omkar Kakade*, Chintan Chitroda, and Vinay Kumar Sankarapu. Dlbacktrace v2: Extending model agnostic interpretability for llms and moes with cuda acceleration. (* - equal contribution) Pre-Print, 2025. URL: https://github.com/AryaXAI/DLBacktrace.

[31] Aditya Tanna, **Pratinav Seth**, Mohamed Bouadi, Utsav Avaiya, and Vinay Kumar Sankarapu. Tabtune: A unified library for inference and fine-tuning tabular foundation models (demo). In ***The Web Conference (WWW) 2026, ACM***, 2026. URL: https://arxiv.org/abs/2511.02802, arXiv:2511.02802.

[32] **Pratinav Seth**, Yashwardhan Rathore, Neeraj Kumar Singh, Chintan Chitroda, and Vinay Kumar Sankarapu. xai_evals : A framework for evaluating post-hoc local explanation methods. Pre-Print, 2025. URL: https://arxiv.org/abs/2502.03014, arXiv:2502.03014.

Position Papers

[33] **Pratinav Seth** and Vinay Kumar Sankarapu. Position: Behavioural assurance cannot verify the safety claims governance now demands. Pre-Print, 2026. URL: https://arxiv.org/abs/2605.15164, arXiv:2605.15164.

[34] Aadit Sengupta*, **Pratinav Seth***, and Vinay Kumar Sankarapu. Interpretability as alignment: Making internal understanding a design principle. (* - equal contribution) **The 1st EurIPS Workshop on Private Governance**, 2025. URL: https://arxiv.org/abs/2509.08592, arXiv:2509.08592.

[35] **Pratinav Seth** and Vinay Kumar Sankarapu. Bridging the gap in xai: Why reliable metrics matter for explainability and compliance. **The 1st EurIPS Workshop on Private Governance**, 2025. URL: https://arxiv.org/abs/2502.04695, arXiv:2502.04695.

Pre-Prints and Shared Tasks

[36] Soham Bhattacharjee, Karun Sharma, Vinay Kumar Sankarapu, and **Pratinav Seth**. Provenance-grounded gating and adaptive recovery in synthetic post-training data curation. Under Review, 2026. URL: https://arxiv.org/abs/2606.11127, arXiv:2606.11127.

[37] Karun Sharma, Soham Bhattacharjee, Vinay Kumar Sankarapu, and **Pratinav Seth**. Document-as-function: Verifiable generation of long-form synthetic documents. Under Review, 2026.

[38] Saisab Sadhu, **Pratinav Seth**, and Vinay Kumar Sankarapu. Forgetting that sticks: Quantization-permanent unlearning via circuit attribution. Pre-Print, 2026. URL: https://arxiv.org/abs/2605.15138, arXiv:2605.15138.

[39] **Pratinav Seth**, Hem Gosalia, Aditya Kasliwal, and Vinay Kumar Sankarapu. Faithfulness is not actionability: Component heterogeneity in discovered circuits. Under Review, 2026.

[40] Saisab Sadhu, Aadit Sengupta, Vinay Kumar Sankarapu, and **Pratinav Seth**. What do compliance detectors read? an audit of activation probes and guard models. Under Review, 2026.

[41] **Pratinav Seth** and Vinay Kumar Sankarapu. Self-calibrating weight-arithmetic safety-drift repair. Under Review, 2026.

[42] **Pratinav Seth**, Anshul Kaushal, Saisab Sadhu, and Vinay Kumar Sankarapu. Drift then repair: A controlled cross-paradigm audit of safety in fine-tuned llms. Under Review, 2026.

[43] **Pratinav Seth**. The off-switch failure: When safety-repair evaluation rewards model collapse. Under Review, 2026.

[44] Ananth Eswar, **Pratinav Seth**, Utsav Avaiya, and Vinay Kumar Sankarapu. Faithfulness to refusal: A causal audit of neuron selectors in llms. Under Review, 2026. URL: https://arxiv.org/abs/2607.05355, arXiv:2607.05355.

[45] Hemang Malik, Gaurav Pradeep, and **Pratinav Seth****. Hgp-nlp at biolaysumm: Leveraging lora for lay summarization of biomedical research articles using seq2seq transformers. *(** - Senior Author) ;* ***BioNLP 2024 Workshop, ACL 2024***, 2024. doi:10.18653/v1/2024.bionlp-1.78.

[46] **Pratinav Seth**, Rashi Goel, Komal Mathur, and Swetha Vemulapalli. Rsm-nlp at blp-2023 task 2: Bangla sentiment analysis using weighted and majority voted fine-tuned transformers. In ***Proceedings of the 1st Workshop on Bangla Language Processing (BLP 2023)***, Singapore, December 2023. ACL. URL: https://arxiv.org/abs/2310.14261, doi:10.18653/v1/2023.banglalp-1.40.

[47] Sriya Rallabandi, Sanchit Singhal, and **Pratinav Seth****. SSS at SemEval-2023 task 10: Explainable detection of online sexism using majority voted fine-tuned transformers. In ***Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), ACL 2023***. (** - Senior Author) ; **Association for Computational Linguistics**. URL: https://arxiv.org/abs/2304.03518, doi:10.18653/v1/2023.semeval-1.171.

[48] **Pratinav Seth***, Akshat Bhandari*, and Kumud Lakara*. Analyzing effects of fake training data on the performance of deep learning systems. *(* - equal contribution) Pre-Print*, 2023. URL: https://arxiv.org/abs/2303.01268.

Selected Blog Posts

[49] Chirag Chawla, Zera Lyngkhoi, **Pratinav Seth**, Utsav Avaiya, Soham Bhattacharjee, Mykola Khandoga, Rui Yuan, and Vinay Kumar Sankarapu. Retail banking case study: Why specialists beat generalists on template-strict workflows. Lexsi Labs, 2026. URL: https://lexsi.ai/resources/articles/retail-banking-case-study-why-specialists-beat-generalists-on-template-strict-workflows-wpx7b.

[50] Chirag Chawla, Zera Lyngkhoi, **Pratinav Seth**, Utsav Avaiya, Soham Bhattacharjee, Mykola Khandoga, Rui Yuan, and Vinay Kumar Sankarapu. The specialization dividend: Aligning a 4b model for wealth management using aligntune. Lexsi Labs, 2026. URL: https://lexsi.ai/resources/articles/the-specialization-dividend-aligning-a-4b-model-for-wealth-management-using-aligntune.

[51] Aditya Tanna, **Pratinav Seth**, Mohamed Bouadi, Utsav Avaiya, and Vinay Kumar Sankarapu. Now shipping: Tabtune regression for tabular foundation models. Lexsi Labs, 2026. URL: https://lexsi.ai/resources/articles/now-shipping-tabtune-regression-for-tabular-foundational-models.

Research Collaborations & Early Experience

Independent Research Collaborations 2024 – 2026
Academic collaborators across IIT Kharagpur, IISc, and MAHE *Remote*
Lesion-aware MAE for glaucoma detection [6]; and introspective LLM detection of multi-person mental manipulation [5].

Research Intern May 2022 – Dec. 2023
KLIV Research Group (PI: Dr. Debdoot Sheet; Mentor: Rakshith Satish), IIT Kharagpur *Remote*
Healthcare AI: Developed an attention-driven dynamic GCN for noisy, multi-label, comorbidity-aware chest radiograph screening; investigated sanity checks for class activation maps in multi-label chest X-ray classification.

Undergraduate Researcher (Mentor: Dr. Abhilash Pai) Apr. 2023 – Dec. 2023
Dept. of Data Science & Computer Application, Manipal Institute of Technology, MAHE *Manipal, India*
Medical AI / Fairness: Studied the impact of pre-training on skin-tone bias in skin-lesion classification (**WACV 2024** [18]), funded by the MAHE UG Research Grant.

Undergraduate Research Assistant (Mentors: Dr. Vidya Rao & Dr. Poornima P.K.) Jun. 2022 – Sep. 2022
Dept. of Data Science & Computer Application, Manipal Institute of Technology, MAHE *Manipal, India*
Cybersecurity & AI: Multi-class malware classification at the intersection of cybersecurity and AI.

Research Collaborator Oct. 2022 – Mar. 2023
Dr. Vijay Kumar BR, NEC Labs & IIT Roorkee *Remote*
Deep metric learning with self-supervised and contrastive learning; vision-based attention models and hyperbolic representation learning.

Research Collaborator Dec. 2023 – Jan. 2024
Dr. Amit Agarwal, Wells Fargo *Remote*
AgriLLM (**NLP4PI @ EMNLP 2024** [17]): seq2seq LLMs (BART, T5, Flan-T5) for agricultural queries from Indian farmers on a highly noisy real-world dataset.

Machine Learning Intern Mar. 2022 – May 2022
Eedge.ai (Talent-Design Platform) *Remote*
Synthetic Data & Interpretability: Built generative models for synthetic tabular data to fine-tune downstream models on a deep-learning talent-design platform; analyzed the effect of synthetic data on the interpretability of the fine-tuned models. Mentor: Mr. Ananta Mahapatra (CTO & Co-Founder).

Data Science (NLP) Intern Jan. 2022 – Feb. 2022
CUREYA (Aspexx Health Solutions Private Limited), Healthcare AI *Remote*
Built NLP pipelines for Reyana, a conversational healthcare chatbot: text classification, summarization, BERT models, and LSTM-based NLG for clinical dialogue.

Projects

- **SNAP Recommendation Engine** (Jul. 2022) **Github**
  A Graph-Based Recommendation Engine that recommends appropriate similar products, co-purchased products, and high-confidence products similar to one viewed by the user using the Amazon SNAP Co-Purchasing Dataset. Used NetworkX, Dask, and Pandas for the back end and Streamlit for the front end for the demo.
- **Prompted Object Detection, Pose Changing & Inpainting** (Mar. 2024) **Github**
  Developed a sequential pipeline using OwlViT and SAM to detect and segment a specific object. Integrated the Zero123 diffusion model to adjust the object's pose based on given coordinates. Finally, employed Stable Diffusion 2 to inpaint the image with the new pose object and background.
- **Customer Segmentation and Visualization using RFM Analysis** (2022) **Kaggle**
  Clustered customers into distinct groups and visualized the results using RFM analysis, K-Means, and Agglomerative Clustering.

Skills Summary

Interpretability: Mechanistic interpretability, circuit analysis & attribution; TransformerLens, Captum, SHAP, LIME, Integrated Gradients, LRP, DL-Backtrace

Post-Training & Alignment: RLHF, DPO, Constitutional AI; PEFT (LoRA, QLoRA, DoRA), TRL (SFTTrainer, DPOTrainer), Axolotl, Unsloth, LlamaFactory, HF Accelerate

LLM Systems & Efficiency: Flash Attention 2, MoE architectures, quantization (bitsandbytes, AutoGPTQ), pruning, ONNX, TensorRT, lm-evaluation-harness
Deep Learning Systems: CUDA programming, Triton kernels, distributed training (DeepSpeed, FSDP), mixed precision; GPUs: H200, H100, RTX 6000 Pro (multi-GPU)
Tabular & Classical ML: XGBoost, LightGBM, CatBoost, TabPFN, Optuna, Ray Tune
AI & Coding Agents: LangChain, LangGraph, Claude Code, Codex, Cursor, Windsurf
MLOps, Serving & Infra: Weights & Biases, MLflow, Ray, Docker, Slurm, Git, Linux, HPC, HF Hub/Datasets, FAISS, AWS/GCP; vLLM, FastAPI, Flask, Gradio, Streamlit, Ollama
Languages & Core Libraries: Python, C++, SQL, Java, C, LaTeX; PyTorch, TensorFlow, Keras, Scikit-Learn, NumPy, Pandas, Seaborn, Matplotlib, OpenCV, PIL, Geopandas, Shapely, NLTK, SpaCy
MOOC: Deep Learning Specialization, **DeepLearning.ai**; 6th Summer School on AI, **CVIT IIITH**

Academic Service

- **Mentor** at AAAI Undergraduate Consortium 2026
- **Reviewer (Major Conferences)**: AAAI 2027 (Program Committee), NeurIPS 2026, ACM AIES 2026, CVPR (2025-26), ECCV (2024-26), ICCV (2025), WACV (2026-27), BMVC 2026, IJCNN (2025)
- **Reviewer (Workshops & Tracks, 2026)**: Mech. Interp. @ ICML, AI for Good @ ICML, TAIGR @ ICML, FMSD @ ICML, FinDS @ ACM SIGMOD, Advances in Financial AI @ ICLR, BlackboxNLP @ EMNLP, NLP4PI @ EMNLP, System Demonstrations @ EMNLP, Actionable Interpretability @ COLM
- **Reviewer (Workshops, 2021–25)**: Actionable Interpretability @ ICML 2025, RegML @ NeurIPS 2025, NLP4PI @ EMNLP 2024, Bayesian Decision-making @ NeurIPS 2024, Frontiers in Probabilistic Inference @ ICLR 2025, SyntheticData4ML @ NeurIPS 2022-23, Fairness-AIMI @ MICCAI 2024, Domain Adaptation @ MICCAI 2023, Topological/Algebraic/Geometric P.R.A. @ CVPR 2023

Education

Bachelor of Technology in Data Science & Engineering — Manipal, India
Manipal Institute of Technology, India; CGPA: 8.31/10 — Oct. 2020 – Jul. 2024

Achievements

- **Undergraduate Research Grant** (Jan. 2023) worth 10,000 INR from Manipal Academy of Higher Education; project *Explainable & Trustworthy Skin Lesion Classification*, leading to a publication at **WACV 2024**.
- Placed **5th of 134** teams, 13th International Cyber Security Data Mining Competition, Nov. 2022.
- Top 10 of 1000 submissions, Bajaj Finserv HackRx 3.0 Hackathon, Jun. 2022.

Invited Talks

AryaXAI Alignment Labs — Jul. 2024 – Present
Research and Product based Talk — *Webinar*

- **Webinar**: Inside the Black Box: Interpreting LLMs with DL-Backtrace (DLB)
- **Webinar**: Beyond Explainability: Evaluating XAI Methods with Confidence Using xai_evals
- **Paper Podcast**: Interpretability Aware Pruning in Medical Imagery

SSI Club (AI Paper-Fest / Paper Presentation Week) — Oct. 2024
Talk: DL-Backtrace by AryaXAI — *Online*

ACM-W Manipal Chapter — Mar. 2024
Introduction to Research — *Manipal, India*

The Data Alchemists, MIT Manipal — Feb. 2024
Data Dialogue — *Manipal, India*

ACM-W Manipal Chapter — Mar. 2023
Research as an Undergraduate — *Manipal, India*

Positions of Responsibility

Mars Rover Manipal — *Jun. 2021 – Dec. 2023*
Researcher — *Manipal, India*

- **Progression**: Rose from Student Trainee to Senior Researcher leading the AI Research Wing, driving the team's technical roadmap across computer vision, multimodal AI, and medical imaging.
- **Recruitment & Training**: Spearheaded recruitment that raised applicant quality; designed and ran training programs for incoming ML researchers.
- **Mentorship**: Mentored 10+ undergraduates (12+ total collaborators) under Dr. Ujjwal Verma; co-authored their publications, including workshop papers at **NeurIPS, CVPR, ICLR, and MICCAI** [19, 20, 21, 22, 23, 24, 25], an AAAI student abstract [7], and a **Nature Scientific Reports** journal paper [8].

The Data Alchemists — *Nov. 2022 – Sep. 2023*
Co-founder & Head of Artificial Intelligence and Machine Learning — *Manipal, India*

- **Co-founding**: Founded and led the AI/ML club, growing it to 30+ members and establishing a lasting community of practice; ran workshops and technical events on ML fundamentals and projects.

Research Society MIT Manipal *Aug. 2022 – Sep. 2023*
Co-President & Mentor AI Research *Manipal, India*

- **Organization**: Led a 90+ member undergraduate research organization across 10+ technical domains; managed recruitment from 250+ applicants, ran member-development pathways, and championed undergraduate research culture. Mentored 10+ undergraduates and co-authored the RSM-NLP team's shared-task papers [45, 46, 47, 48].

Conferences Attended

ICML 2026 (Seoul, South Korea); The Web Conference (WWW) 2026 (Dubai, UAE; attended remotely); EurIPS 2025 (Copenhagen, Denmark); MICCAI 2025 Workshop (Daejeon, South Korea); ICML 2025 (Vancouver, Canada); ICLR 2025 (Singapore); AAAI 2023 (Washington, D.C., USA).