Pratinav Seth

seth.pratinav@gmail.com · pratinavseth.github.io · Google Scholar · GitHub · LinkedIn

Research Summary

Research scientist and technical lead working on post-training, mechanistic interpretability, and safety for foundation models, with research spanning LLMs, tabular foundation models, and vision. Leads the Model Science group at Lexsi Labs, building research systems from alignment and interpretability tooling to foundation-model training and evaluation. Published at ICML, ACL, WWW, MIDL, and Nature Scientific Reports, alongside NeurIPS, ICLR, CVPR, MICCAI, and ICML workshops.

Research Experience

Lead Research Scientist (Present); Research Scientist (Jul. 2024 – Mar. 2026) Jul. 2024 – Present
Lexsi Labs / AryaXAI Alignment Labs (Aurionpro Solutions Group) *Remote*

- **Research Leadership**: Direct research across interpretability, alignment, safety, tabular foundation models, and agents; scaled the team to 20+ researchers and interns across India and Paris, owning hiring, roadmap, grants, mentoring, and research-to-product integration while remaining technically hands-on.
- **Post-Training & Alignment**: Architected and scaled AlignTune (37⋆) into a production-grade, multi-backend post-training system (SFT, preference optimization/RL, PEFT, model merging); built CuratorKIT (24⋆), a provenance-grounded data-curation and synthetic-generation pipeline, with research on verifiable long-form synthetic document generation.
- **Mechanistic Interpretability**: Built CircuitKIT (14⋆), a circuit discovery, evaluation, and application toolkit; co-developed C-$\Delta\theta$, a circuit-restricted weight-arithmetic method for selective refusal (**ICML 2026 Workshop**); ongoing research on circuit attribution and circuit faithfulness.
- **LLM Safety**: Co-built SafeTune (7⋆), a unified library for auditing and repairing safety drift in fine-tuned LLMs; built an inference-time alignment-transfer method (**ICML 2026 Workshop**); research program spans safety-drift repair, safety-repair evaluation, and auditing of compliance detectors and guard models.
- **Tabular Foundation Models**: Led problem framing, architecture design, and benchmarking for the ORION tabular-foundation-model series – Orion-Bix (**WWW 2026**) and Orion-MSP (AITD Workshop, EurIPS 2025); architected TabTune (116⋆, **WWW 2026** Demo), an inference/fine-tuning toolkit for tabular foundation models, with follow-on research in distillation, ensembling, structured health data, and credit-risk prediction.
- **Interpretability, Evaluation & Model Systems**: Rearchitected DLBacktrace v2 (26⋆) for CUDA-accelerated, model-agnostic interpretability across LLMs and MoEs; built xai_evals (15⋆), a benchmarking framework for explainability methods, and an internal LLM evaluation library; lead structured-pruning and quantization work for efficient model deployment.

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

- **Geospatial Climate AI**: Bachelor's thesis project (first-author, **ICML 2025** & ICLR 2025 ClimateAI Workshop): built the Alberta Wells Dataset, a novel satellite-imagery dataset for detecting abandoned oil and gas wells, working with domain experts; benchmarked deep-learning models for geospatial climate applications.
- **Mentor**: Dr. David Rolnick (McGill University, Université de Montréal, Mila).

Computer Vision Research Intern Jun. 2023 – Oct. 2023
Bosch Corporate Research *Bangalore, India*

- **Generative Data Augmentation**: Developed a generative data-augmentation pipeline for safety-critical autonomous driving perception; used latent diffusion models to synthesize hard-negative samples, improving CV fine-tuning.

Selected Research

48 papers overall; 31 peer-reviewed · 210+ citations · h-index 8 · Publicly Available list: Google Scholar.

LLM Alignment, Safety & Mechanistic Interpretability

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

[2] 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.

[3] 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***. URL: https://arxiv.org/abs/2505.20679.

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

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

[6] 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.

[7] 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.

[8] Vinay Kumar Sankarapu, Chintan Chitroda, Yashwardhan Rathore, Neeraj K. Singh, and **Pratinav Seth**. Dl-backtrace: A model agnostic explainability for any deep learning models. *IJCNN*, 2025. URL: https://arxiv.org/abs/2411.12643.

Tabular Foundation Models

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

Vision, Medical AI & Climate

[14] **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.

[15] 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.

[16] 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.

[17] 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.

Academic Service & Recognition

Reviewer: AAAI 2027 (Program Committee), NeurIPS 2026, CVPR, ECCV, ACL/EMNLP, WWW, and other major venues (2024–2027). **AAAI Undergraduate Consortium**: Scholar (2023) → Mentor (2026).

Education

B.Tech. in Data Science & Engineering – Manipal Institute of Technology, India; CGPA 8.31/10 (Oct. 2020 – Jul. 2024)

Technical Skills

LLM Post-Training: SFT, preference optimization, RLHF, DPO, PEFT/LoRA, TRL, Axolotl, Unsloth, LlamaFactory
Interpretability & Safety: Mechanistic interpretability, circuit analysis, TransformerLens, LRP, IG, safety evaluation
Foundation Models & Systems: PyTorch, CUDA, MoE, vLLM, quantization, pruning
Infrastructure: Single and multi-GPU handling, Slurm, Docker, Hugging Face, W&B, MLflow, AWS/GCP
Languages: Python, C++