Varad Pimpalkhute

(425) 543 7285 pimpalkhutevarad@gmail.com Website Google Scholar LinkedIn

Work Experience

Institute of Foundational Models **Jun 2025 – Present**

Research Scientist, Post-Training RL (Full time) *Sunnyvale, California*

- Leading frontier post-training RL research for reasoning, SWE, terminal, and deep-research agents, spanning long-horizon tool use, verifiable reward environments, trajectory generation, and off-policy optimization.
- Built and scaled asynchronous RL training infrastructure, improving throughput by 4× and stabilizing RL post-training across 400B MoE, 70B dense, and smaller dense/MoE models on 1000s of GPUs.
- Leading IFM's agentic RL infrastructure for SWE-bench-style and terminal-based tasks, including coding sandboxes, rollout orchestration, training workflows, reward environments, and scalable evaluation pipelines.
- Led coding-agent task development and post-training for K2 Horizon, owning SWE and terminal environments, rollout/evaluation pipelines, and agentic training workflows used to train and evaluate the released models. (Release)
- Researching off-policy RL scaling laws for LLM post-training, including sampling-compute allocation, IsoCompute analysis, data reuse, critical batch size, and compute-optimal reasoning model training.
- Led and mentored interns and researchers across RL post-training projects, including IsoCompute Playbook, agentic task training, deep-research agents, and SWE/terminal environments.
- Co-led K2-Think-V2 technical blog and model release efforts, contributing to a sovereign reasoning model with state-of-the-art open-source reasoning performance on public leaderboards. (Blog)

Articul8.ai **Jun 2024 – May 2025**

Applied AI Researcher (Full time) *Seattle, Washington*

- Led end-to-end training of an 8B Verilog code model for code generation, completion, debugging, and unittest tasks.
- Designed a memory manager for handling short-term and long-term memory across sessions. Introduced hybrid reasoning trajectories in place of naive chain-of-thought (CoT), boosting performance on financial reasoning datasets by 7%.
- Pretrained on 40B tokens, post-trained with SFT & GRPO using verifiable rewards, achieving a 2× improvement over open-source baselines. (Media Article, Blog)
- Led efforts in synthetic dataset generation and full-supervised fine-tuning of LLMs improving task performance across agentic benchmarks.
- Automated evaluation pipelines for the agentic framework, leveraging LoRA optimizations to fine-tune function callers on a synthetic dataset, achieving results comparable to GPT-4o on Berkeley function calling. (Slide Deck)
- Designed and benchmarked a tool to optimize model serving across TGI and vLLM, considering user constraints and hardware availability.

Allen Institute for Artificial Intelligence (AI2) **Jan 2024 – May 2024**

Graduate Student Researcher (Part time) *Seattle, Washington*

- Wrote scripts for multi-node training LLMs utilizing 4D (tensor, pipeline, context, data) parallelism. Additionally, adopted quantization techniques to migrate training to a lower compute regime.
- Developed an RL-based method, RL4F, for automatic schema generation, enhancing explainability in moral reasoning tasks by transitioning from reliance on manual schemas to automated schemas in LLM such as OLMO, LLaMA 2.

Amazon Science **May 2023 – Dec 2023**

Applied Scientist Intern (Full time) *Boston, Massachusetts*

- Developed a query encoder that incorporates knowledge from an Large Language Model, achieving a 2.8% improvement in downstream retrieval performance without increasing latency.
- Created a framework for the automatic evaluation of failure modes in Retrieval Augmented Generation (RAG) pipelines, leading to a 40% reduction in error rates and improving the reliability of generated outputs.

Tata Consultancy Services **May 2020 – Jul 2022**

Systems Engineer (Researcher) (Full time) *Mumbai, India*

- Enhanced the algorithm design of the gradient-based family of meta-learning algorithms, achieving a 3% increase in generalization accuracy for multi-distribution task setups on few-shot computer vision datasets.
- Proposed generation of efficient subnetworks in the architecture that resulted in significant gains in accuracy and speed up of more than 2x on various few-shot vision benchmark datasets using distributed computing.

Cedars Sinai (Computational Biomedicine Lab) **May 2023 – May 2024**

RADS Scholar (Part time) *Los Angeles, California*

- Designed an automatic pipeline for identification of medication change by leveraging LLMs for topic discovery.
- Developed a framework that predicts the home location of a Twitter user to the nearest location. Framework was used for tracking epidemics and infectious diseases, and improvement by 5x over the previous database.

Publications

Recent publications, complete list here

- **Unlocking Lossless Speedups in LLMs via Discrete Diffusion**, *arXiv preprint 2026* **[PDF]**
  Developed diffusion-augmented LLMs for lossless parallel token generation, achieving up to 3× inference speedups while preserving the underlying autoregressive model distribution.
- **Efficient Agentic Reasoning Through Self-Regulated Simulative Planning**, *arXiv preprint 2026* **[PDF]**
  Developed a self-regulated agentic reasoning framework that adaptively decides when and how deeply to plan, improving reasoning efficiency through learned simulative planning.
- **IsoCompute Playbook: Optimally Scaling Sampling Compute for LLM RL**, *arXiv preprint 2026* **[PDF]**
  Developed scaling insights for allocating sampling compute in RL post-training, providing practical guidance for compute-optimal LLM reinforcement learning.
- **DocQIR-Emb: Document Image Retrieval with Multi-lingual Question Query**, *UniReps Workshop 2025*
  Built a multilingual question-query benchmark for document image retrieval, evaluating representation learning methods for cross-lingual visual-document understanding.
- **K2-V2: A 360-Open, Reasoning-Enhanced LLM**, *arXiv preprint 2025* **[PDF]**
  Contributed to the development of a fully open reasoning-enhanced LLM, covering post-training, evaluation, and reasoning capability improvements.
- **K2-think: A parameter-efficient reasoning system**, *arXiv preprint 2025* **[PDF]**
  Developed a 32B parameter-efficient reasoning system, outperforming larger open-weight models across reasoning benchmarks through optimized post-training and inference.
- **Bottleneck Discovery: Towards interpretable schemas using RLHF**, *in review*
  Learned conceptual bottleneck using PPO for aligning LLMs on the task of morality.
- **Relational knowledge discovery using in-context learning**, *in review*
  We analyze the ability of LLMs to discover FOL queries in a real-world KG setting.
- **Soft Query Expansion: Learned Representations of Queries Expanded by LLMs**, *ECIR 2024* **[PDF]**
  Improved on DPR by incorporating knowledge from LLMs resulting in a 2.8% improvement.
- **GEMS: Generating Efficient Meta Subnets**, *WACV 2023* **[PDF]**
  Addressed the challenge of loss of generalization on MAML in a multi-distributional setting on computer vision datasets.
- **Transformer models for classification on COVID19 Twitter posts**, *WNUT Workshop at EMNLP 2020* **[PDF]**
  Built a system to identify informative COVID19 posts on Twitter for use in downstream applications. The system was ranked 13^{th} in WNUT shared task.

Education

University of Massachusetts Amherst **Sep 2022 – May 2024**
Masters of Science in Computer Science (GPA: 4.00/4.00) *Amherst, MA*

Indian Institute of Information Technology, Nagpur **Jul 2017 – May 2021**
Bachelor of Technology in Electronics & Communication Engineering (GPA: 3.68/4.00) *Nagpur, India*

Skills

Large Language Models: LangChain, LlamaIndex, Faiss, Chroma
Programming Languages: C++/C, Python, HDL, Bash, MATLAB, Raspberry Pi, LaTeX, Git
Machine Learning: PyTorch, AllenNLP Flair, PyTorch Lightning, Hugging Face
Data Processing: Pandas, Numpy, NLTK, Matplotlib
Miscellaneous: AWS (Sagemaker, Lambda, EC2), HTML, CSS, Linux, CentOS, Streamlit

Academic Experience

- **IESL, UMass Amherst** **Guide:** Professor Andrew McCallum
  Investigated the use of box embeddings as an usecase for hierarchical representation, and further replace vector encoders with box-like encoders. (Investigations)
- **Fatima Fellowship** **Guide:** Dr. Maruan Al-Shedivat
  Investigating Gradient Hacking (GH) methods to reduce the negative-transfer (interference) in multi-task training. Analyzing ways to propose a probabilistic framework for the GH methods.

Projects

- **Load Test Framework**, A tool for load testing and performance benchmarking. Load Test Framework is a tool designed for load testing and performance benchmarking of software systems. It helps in identifying bottlenecks and optimizing system performance by simulating various load conditions. **Code Link:** GitHub Link

- **Financial Reports Chatbot**, An AI-driven chatbot for financial report analysis Financial Reports Chatbot is an AI-powered chatbot designed to analyze and summarize financial reports. It utilizes natural language processing techniques to provide insights and answer queries about financial data. **Code Link:** GitHub Link

- **Function Call Generation, and Hallucination Mitigation.** Developed an arbitration mechanism for balancing function calls and parametric knowledge in LLM responses. Implemented techniques to detect and reduce hallucinations using confidence thresholding and zero-shot prompting. **Report:** PDF

- **Metabolts**, A library for high-end research in Meta learning **Guide:** Sébastien Arnold
  Metabolts is a software library for meta-learning research. It builds on top of PyTorch Lightning to accelerate two aspects of meta-learning research cycle: fast-prototyping and correct reproducibility. **Code Link:** GitHub Link

- **Through the Bias-oscope: A Text2Image Generation Journey** **Guide:** Subhransu Maji
  Analysed inherent biases present in two state-of-the-art text-to-image generation models, Stable Diffusion and DALL-E2. Proposed methods for mitigating the bias present in sampled LION-400M dataset.

- **MetaFaaS: learning-to-learn on serverless** **[PDF]**
  Accelerated training of meta-learning architectures by ∼9x on AWS Lambda instances in a distributed setting. The work was published at Big Data in Emergent Distributed Env. at SIGMOID 2022.