

Harsh Kaushik

harshkaushik11000@gmail.com | +91 8287108466 | [GitHub](#) | [LinkedIn](#) | [Portfolio](#) | [Hugging Face](#)

Professional Summary

AI Systems & Machine Learning student focused on efficient and affordable LLM inference, training Small Language Models (SLMs), and reproducing state-of-the-art research. Experienced in building language models from scratch, post-training open-source LLMs, and developing large-scale training datasets for efficient AI systems.

Skills

- **Programming:** Python • C++ • C
- **Deep Learning & ML:** PyTorch • Transformers • NumPy • Pandas • Scikit-learn
- **LLM Ecosystem:** Hugging Face • vLLM • llama.cpp • Ollama • LM Studio • SentencePiece • Tokenizers
- **Model Training:** Pre-training • Continued Pre-training (CPT) • Supervised Fine-tuning (SFT) • LoRA • QLoRA • PEFT • Tokenizer Training • Mixed Precision
- **Data Engineering:** Parquet • Dataset Preprocessing • Data Cleaning • Multiprocessing
- **AI Systems:** Efficient LLM Inference • Transformer Architectures • Small Language Models (SLMs) • Model Compression • Mixture of Experts (MoE)
- **Development:** Linux • Git • VS Code • Jupyter • Kaggle • Google Colab

Education

University School of Automation and Robotics (USAR) | Bachelor of Technology (Data Science) | 2025–2029

Projects

- **TinyStories-17M** [Completed] | Trained a 17.2M-parameter decoder-only Transformer language model from scratch using a 2.15M-sample TinyStories-style dataset. [[Hugging Face](#)]
- **TurboLLM** [Ongoing] | Research project exploring efficient LLM inference on consumer GPUs through throughput benchmarking, memory optimization, and decoding strategies. [[GitHub](#)]
- **LLM Post-training** [Completed] | Fine-tuned open-source models for math, legal, and programming tasks, publishing several models on Hugging Face. [[Hugging Face](#)]
- **Custom LLM Training Datasets** [Completed] | Developed instruction-tuning datasets and preprocessing pipelines for reasoning and programming tasks, publishing them on Hugging Face. [[Hugging Face](#)]

Publications & Open Source

- Published multiple open-source language models on Hugging Face.
- Published multiple instruction-tuning datasets on Hugging Face.
- Maintains open-source repositories focused on LLM training, post-training, and efficient inference.

Research Reading

[Attention Is All You Need] [TinyStories] [SmoLLM2] [Phi-1] [Phi-1.5] [Scaling Laws] [Chinchilla] [DeepSeek-V3] [LoRA] [InstructGPT]